Problems of Computational Mathematics and Mathematical Modelling

Edited by G.I. MARCHUK and V.P. DYMNIKOV

Advances in Science and Technology in the USSR

Mathematics and Mechanics Series

MIR Publishers
Moscow

Problems of Computational Mathematics and Mathematical Modelling

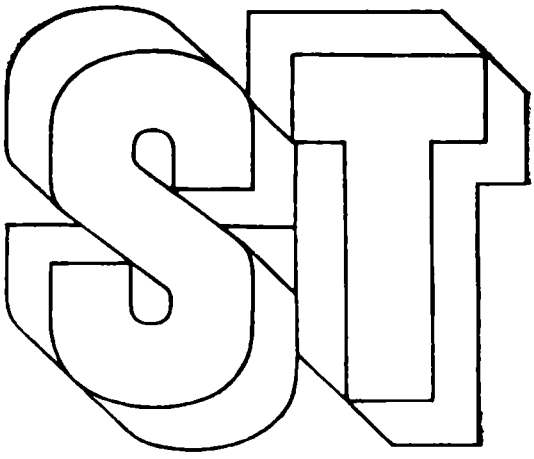
Edited by G.I. MARCHUK, Mem. USSR Acad. Sc., and V.P. DYMNIKOV, D.Sc. (Phys.-Math.)

The work presented in this collection covers three main topics. The first deals with solving the systems of linear algebraic equations that result from the discretization of mathematical physics problems by finite-difference and projection-difference methods. Another is associated with the mathematical modelling of general atmospheric and oceanic circulation (in particular, consideration is given to optimization models for environmental protection). The third topic deals with the mathematical modelling of infectious diseases. Some papers are concerned with neutron transport equations, vector algorithm, and randomization of Monte Carlo methods. These papers develop the basic theory of vector stochastic algorithms for solving integral equations and of "double randomization" algorithms. The final paper discusses the technology of software application packages.

Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Mir Publishers Moscow

Advances
in
Science
and
Technology
in
the USSR



Mathematics and
Mechanics Series

Problems of Computational Mathematics and Mathematical Modelling

Edited by
G. I. MARCHUK,
Mem. USSR Acad. Sc.,
and V. P. DYMNIKOV,
D. Sc. (Phys-Math.)

Translated from the Russian by
Vladimir Shokurov

Mir Publishers
Moscow

Проблемы вычислительной математики
и математического моделирования
Под редакцией академика *Г. И. Марчука*
и д-ра физ.-мат. наук *В. П. Дымникова*
Издательство «Мир» Москва

First published 1985

На английском языке

© Издательство «Мир», 1985
© English translation, Mir Publishers, 1985

Contents

Preface 6

1 A Generalized Conjugate Gradient Method in a Subspace and Its Applications *by G. I. Marchuk and Yu. A. Kuznetsov* 10

2 Problems of Projection-Grid Methods in Transport Theory *by G. I. Marchuk and V. I. Agoshkov* 33

3 Variational Forms of Problems in Transport Theory *by V. I. Agoshkov and V. V. Smelov* 58

4 Vector Algorithms and Randomization of Monte Carlo Methods *by B. A. Kargin, G. . Mikhailov, M. A. Nazaraliev, and K. K. Sabelfeld* 84

5 Numerical Simulation of a Large-Scale Atmospheric and Oceanic Circulation *by G. I. Marchuk, V. P. Dymnikov, V. B. Zalesny, V. N. Lykosov, and V. Ya. Galin* 124

6 Methods of Mathematical Modelling of Sea and Oceanic Processes *by A. S. Sarkisyan, V. P. Kochergin, V. I. Klimok, A. A. Kordzadze, V. I. Kuzin, V. A. Sukhorukov, and A. V. Schcherbakov* 167

7 Optimization Models for Environmental Protection *by V. V. Penenko, A. E. Aloyan, N. N. Obraztsov, A. V. Panarin, A. V. Protasov, and V. F. Raputa* 198

8 Mathematical Modelling of Infectious Diseases *by G. I. Marchuk, L. N. Belykh, and S. M. Zuev* 223

9 Software for Designing Electrophysical Devices *by V. P. Ilyin, A. V. Gavrilin, B. I. Golubtsov, N. I. Gorbenko, V. A. Kateshov, G. S. Popova, V. M. Sveshnikov, A. L. Urvantsev and M. V. Urev.* 241

Preface

The rapid development of computer science is heavily influencing the rate of scientific and technological progress. Many of what were once called qualitative sciences, such as biology, geography, geology, ecology and medicine, are becoming more and more quantitative, while the traditionally creative spheres of human activity—design and technology—are now aided by interactive computing. Moreover, optimal solutions have become possible in various areas of human knowledge.

The mathematical modelling of complex processes was the first subject to experience the impact of the developments in computer technology because it became possible to use powerful mathematical methods that had previously been infeasible. The theories of identification, optimization, and dynamic programming, for example, had already been developed and employed, but were not widely used.

Numerical methods developed in close connection with developing computer technology. Moreover, each level in the development of computer technology influenced computational mathematics. Take, for example, variational methods. They were very important in pre-computer mathematics as it was possible to state a problem of mathematical physics or engineering as a problem on the minimum of a functional and construct an approximate solution in the form of a small but informative set of trial functions. In many cases the success of the solution depended on the choice of the simple functions and this constituted the creative component of the study. The advent of electronic computers and the rapid development of finite-difference methods for solving problems rendered variational methods almost obsolete until even more powerful computers appeared. These made it possible to pass from difference approximations to variational-difference ones, presenting science for a second time with the finite-element method and the method of spline functions, the origins of both going back to pre-computer mathematics. As a result of this loop a powerful set of difference methods has become available, enriched with the idea of a best, in a certain sense, approximation of the original operators of problems without violating their definiteness. These methods are now used everywhere.

The selection of the most efficient algorithms for the numerical solution of a problem was also interesting. At first the quality of the

algorithm, that is its ability to produce a solution within the required accuracy given a comparatively low computer speed and working storage, was extremely important. During that time a huge number of numerical methods were investigated. Methods for splitting complicated problems into simple ones, schemes for continuous computation, iterative processes with optimization, etc. all appeared. What is important is that in order to select algorithms they had to be subjected to deep theoretical analysis. This promoted the development of theoretical computational mathematics and, as a consequence, a class of efficient algorithms has been identified that have ever since formed a firm basis for computational mathematics.

Further progress in the capacity of computers gave rise for a while to some complacency among investigators, who preferred algorithms that are simple but take up a great deal of computer time to good quality algorithms which are more difficult to use. Intellectual work was being saved at the expense of being less economic in the solution of a problem. Naturally, this situation could not last for long since practical needs required the limits of computers possibilities: first two-dimensional problems were tackled, then three-dimensional ones were. Once again interest in good quality algorithms grew. Investigators passed from splitting the operators of problems into simpler ones to decomposing domains, and large problems began to be reduced to sets of problems that are small and that can easily be implemented on modern multiprocessor computers that work in parallel. Generally speaking, we are already in the era of parallel computation, and this trend will undoubtedly dominate in the foreseeable future. For computational mathematics this means a renewed interest in the methods of splitting problems according to operators and decomposing according to domains.

When speaking about the rising interest in good quality algorithms, it should be said that it too is due to the use of new computing technology, such as pipeline processors, and fast special-purpose processors that can solve an auxiliary problem within the time required for a few cycles of the basic processor. Examples of the latter are vector processors, matrix processors, macropipeline processors, and fast Fourier transform processors. In particular, these processors enable local or global optimization.

The future of computer science lies in the intellectualization of technology and creation of knowledge bases. The next decade will undoubtedly see contributions to the treasury of science, such as optimization methods together with interactive and artificial intelligence systems. As to new methods of computational mathematics we can anticipate a growing interest in the theory of semigroups for solving problems in mathematical physics, a broad approach to the use of optimal processes for implementating variational functionals and Monte Carlo methods.

Nowadays technological potential of computational mathematics based on large-scale integrated circuits is so extensive that the creation of computers is itself becoming a subject of investigation of mathematicians. The time has now come to formulate computer architectures as the mapping of numerical methods and data-base control. To do this will require the development of new directions of computational mathematics and computer technology. The driving force behind the creation of new methods and algorithms is first of all associated with the need to solve ever more complicated problems. Hence applied problems can be justly considered to be the source of new computational mathematics. Indeed good computing algorithms originally arose to solve applied problems, and only then after years of wider applications were they abstracted to form the basis of computational mathematics.

The writer of these lines and his pupils, collaborators, and colleagues have for many years regarded computational and applied mathematics as a single system of knowledge. This has allowed us to make some advances in a number of particular areas of science with a subsequent generalization of the results obtained to computational mathematics. This can be seen from the papers included in this collection.

"A Generalized Conjugate Gradient Method in a Subspace and Its Applications" by Marchuk and Kuznetsov presents a new approach to the generalization of the classical method of conjugate gradients with preconditioning. The approach is based on considering the method in a specially chosen subspace of a vector space rather than in the entire space. The method is applied to the optimization of the iterative methods of fictitious components and domain decomposition in the corresponding subspaces.

"Problems of Projection-Grid Methods in Transport Theory" by Marchuk and Agoshkov and "Variational Forms of Problems in Transport Theory" by Agoshkov and Smelov discuss questions arising in solving a class of problems based on the transport equation. The first paper covers some of the problems when projection-grid methods are used, such as the smoothness of a solution, the choice of basis functions, and the extention of the functions with the smoothness preserved, and suggests solution methods for a class of problems. The second paper formulates a variational statement of a problem in a three-dimensional region with piecewise smooth boundaries for the method of spherical harmonics. An approach to the symmetrization of the original transport equation is also considered.

"Vector Algorithms and Randomization of Monte Carlo Methods" by Kargin *et al.* presents basic conclusions from the theory of vector stochastic algorithms for solving systems of integral equations and "double randomization" algorithms for solving problems with random parameters. It deals with the basic theory and construction of these

algorithms for transport theory taking into account the polarization of light and stochastic heterogeneity of the medium, and with stochastic problems of elasticity theory.

"Numerical Simulation of a Large-Scale Atmospheric and Oceanic Circulation" by Marchuk *et al.*, "Methods of Mathematical Modelling of Sea and Oceanic Processes" by Sarkisyan *et al.*, and "Optimization Models for Environmental Protection" by Penenko *et al.* cover the main computational aspects of modelling atmospheric and oceanic circulation, the simultaneous atmospheric and oceanic circulation, and the dynamics of pollutant propagation in the atmosphere and ocean. These problems are some of the most complicated problems in modern physics and require large computational costs. When implementing these models the efficiency of the computational algorithms is of primary importance. The algorithms presented in the papers are based on splitting the original problem by both physical processes and geometrical parameters. Solution methods for problems with a small parameter of a higher derivative, the procedure for investigating the response of discrete models to variations in the input parameters, and methods of experimental design are considered.

In recent years mathematicians have shown a growing interest in one of the biological sciences, immunology. The rapid development of theoretical and experimental immunology was stimulated by (and stimulated in turn) the application of mathematical modelling. The mathematical modelling of infectious diseases are discussed in the paper by Marchuk, Belykh, and Zuev.

Finally, the paper by Ilyin *et al.* uses the problem of designing electrophysical devices as an example to demonstrate the current approach to constructing computer-aided design systems. As already stated above, this direction is naturally connected with the future of computer science, the intellectualization of technology and creation of data bases.

G. Marchuk

1 A Generalized Conjugate Gradient Method in a Subspace and Its Applications

G. I. Marchuk and Yu. A. Kuznetsov

In recent years much work has appeared devoted to various generalizations and applications of the classical conjugate gradient method first developed in [1.1] and [1.2]. In this paper we shall present our approach to the generalization of the method, proposed in [1.3] and subsequently developed in [1.4]-[1.6]. The basic feature of the approach consists in considering the method as a computational process in a subspace arising as a rule owing to a special choice of initial guess.

The paper consists of five sections. Section 1.1 deals with general problems of the theory of the generalized conjugate gradient method in a subspace. Section 1.2 discusses some simplest specific variants of the method. The remaining three sections are devoted to applications of the method to the optimization, in the corresponding subspaces, of the method of simple iteration for the kinetic transport equation, of the fictitious components method, and the block relaxation method which uses geometrical decomposition of the domain in which the solution of the original differential problem is defined.

1.1. GENERAL THEORY

Consider the system of linear algebraic equations

$$A\mathbf{u} = \mathbf{f} \tag{1.1}$$

with a real quadratic matrix A and a vector $\mathbf{f} \in \operatorname{Im} A = AE_N$, where E_N is the space of N-dimensional real vectors with the usual scalar product $(\cdot, \cdot)$ and the corresponding norm $\| \cdot \| = (\cdot, \cdot)^{1/2}$. Notice that it is not assumed here that the matrix A is nonsingular, but it is required that the original system (1.1) be consistent.

We next introduce matrices D and H and subspaces $U_A \subseteq E_N$, assuming the following:

the subspace U_A is invariant under the matrix $S = AH$;

the matrix D is self-adjoint and positive definite in U_A, i.e. for any $\xi, \eta \in U_A$ the equality $(D\xi, \eta) = (\xi, D\eta)$ holds and $(D\xi, \xi) > 0$ holds for any nonzero $\xi \in U_A$.

The assumption about D allows us to define in U_A a new scalar pro-

duct and a new norm:

$$(\xi, \eta)_D = (D\xi, \eta), \quad \xi, \eta \in U_A,$$
$$\| \xi \|_D = (\xi, \xi)_D^{1/2}, \quad \xi \in U_A.$$

Now we make yet another very important assumption that in U_A the matrix $S = AH$ is D-self-adjoint and D-positive definite, i.e. for any $\xi, \eta \in U_A$ we have $(S\xi, \eta)_D = (\xi, S\eta)_D$ and for any nonzero $\xi \in U_A$ we have $(S\xi, \xi)_D > 0$. Notice that this assumption is equivalent to the requirement that DS be self-adjoint and positive definite in U_A in the sense of the usual scalar product $(\cdot, \cdot)$.

Under the assumptions made we have

Lemma 1.1. *Let $t = \dim U_A$. Then U_A is the span of some D-orthonormal system of eigenvectors $\psi_1, \ldots, \psi_t$ of S which correspond to its positive eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_t$.*

Let some nonzero vector $\xi \in U_A$ and some integer $r \geqslant 1$ be given. Then the span of the system of vectors $S\xi, \ldots, S^r\xi$ is called the Krylov [1.7] space and denoted by

$$\mathcal{K}_r(\xi) = \mathcal{K}(S\xi, \ldots, S^r\xi).$$

It is easy to see that $\mathcal{K}_p(\xi) = \mathcal{K}_r(\xi)$ for any $r > p$, where $p = p(\xi)$ is the maximum whole number for which the system of vectors $S\xi, \ldots, S^p\xi$ is linearly independent. In particular, $\mathcal{K}_r(\xi) = \mathcal{K}_{\hat{t}}(\xi)$ for any $r > \hat{t}$, where $\hat{t}$ is the number of distinct values among the previously determined eigenvalues $\lambda_1, \ldots, \lambda_t$ of S. Note that for any $\xi \in U_A$ we have $p(\xi) \leqslant \hat{t}$.

Now suppose that $\xi^0 = A\mathbf{u}^0 - \mathbf{f} \in U_A$ for some given vector $\mathbf{u}^0 \in E_N$ and that we know a system of vectors $\mathbf{g}_1, \ldots, \mathbf{g}_p$ such that the vectors $\mathbf{w}_k = A\mathbf{g}_k$, $k = \overline{1, p}$, form a D-orthogonal basis of $\mathcal{K}_p(\xi^0)$. Then, using the linear dependence of the vectors $\xi^0, S\xi^0, \ldots, S^p(\xi^0)$ (with $p = p(\xi)$ previously determined) it is easy to see that

$$\xi^0 = \sum_{k=1}^{p} \beta_k \mathbf{w}_k, \quad \beta_k = (\xi^0, \mathbf{w}_k)/\| \mathbf{w}_k \|^2, \quad k = \overline{1, p}, \tag{1.2}$$

and hence the solution $\mathbf{u}^*$ of the original system (1.1) can be found using the computational process

$$\mathbf{u}^k = \mathbf{u}^{k-1} - \beta_k \mathbf{g}_k, \quad k = \overline{1, p}, \tag{1.3}$$

where $\mathbf{u}^* = \mathbf{u}^p$. Let us investigate in detail the problem of constructing the required basis $\mathbf{w}_1, \ldots, \mathbf{w}_p$ for some arbitrary but fixed vector $\xi \in U_A$.

This problem can be solved using the usual Gram-Schmidt D-ortho-

gonalization process:

$$\mathbf{w}_k = \begin{cases} S\xi, & k=1, \\ S^k\xi - \sum\limits_{i=1}^{k-1} a_{k,i}\mathbf{w}_i, & k>1, \end{cases} \tag{1.4}$$

$$a_{k,i} = \frac{(S^k\xi,\ \mathbf{w}_i)_D}{\|\mathbf{w}_i\|_D^2}, \quad i=\overline{1,\ k-1},\ k=\overline{1,\ p}$$

or equivalently the process

$$\mathbf{w}_k = \begin{cases} S\xi, & k=1, \\ S\mathbf{w}_{k-1} - \sum\limits_{i=1}^{k-1} b_{k,i}\mathbf{w}_i, & k>1, \end{cases} \tag{1.5}$$

$$b_{k,i} = \frac{(S\mathbf{w}_{k-1},\ \mathbf{w}_i)_D}{\|\mathbf{w}_i\|_D^2}, \quad i=\overline{1,\ k-1},\ k=\overline{1,\ p}$$

which, by virtue of the easily verifiable equations $b_{k,i} = 0$ for $i = \overline{1,\ k-1}$, can be represented in the three-term form

$$\mathbf{w}_k = \begin{cases} S\xi, & k=1 \\ S\mathbf{w}_1 - \alpha_2\mathbf{w}_1 & k=2 \\ S\mathbf{w}_{k-1} - \alpha_k\mathbf{w}_{k-1} - \gamma_k\mathbf{w}_{k-2}, & k>2, \end{cases} \tag{1.6}$$

$$\alpha_k = \frac{(S\mathbf{w}_{k-1},\ \mathbf{w}_{k-1})_D}{\|\mathbf{w}_{k-1}\|_D^2}, \quad \gamma_k = \frac{(S\mathbf{w}_{k-1},\ \mathbf{w}_{k-2})_D}{\|\mathbf{w}_{k-2}\|_D^2} = \frac{\|\mathbf{w}_{k-1}\|_D^2}{\|\mathbf{w}_{k-2}\|_D^2}, \quad k=\overline{1,\ p}$$

called the Lanczos orthogonalization process [1.8], [1.9]. We show that the orthogonal basis for $\mathscr{K}_p(\xi)$ can be constructed using still more compact formulas, by combining (1.6) with process (1.3) for residual vectors:

$$\xi^k = \xi^{k-1} - \beta_k\mathbf{w}_k, \quad \beta_k = \frac{(\xi^{k-1},\ \mathbf{w}_k)_D}{\|\mathbf{w}_k\|_D^2}, \quad k=\overline{1,\ p}, \tag{1.7}$$

where $\xi^0 = \xi$.

Suppose, taking into account $\beta_1 = \frac{(\xi^0,\ S\xi^0)_D}{\|S\xi^0\|_D^2} > 0$, that $\beta_k \neq 0$ for some $k \geqslant 1$. We show that $\beta_{k+1} \neq 0$ (if $k+1 \leqslant p$). We shall use the fact that the vectors $\mathbf{w}_i$, $i = \overline{1,\ p}$, of process (1.6) are necessarily nonzero, $S\xi^{k-1} \in \mathscr{K}_k(\xi^0)$, and the residual vector ξ^k is D-orthogonal to $\mathbf{w}_i$, $\overline{i=1,\ k}$.

Since $\mathbf{w}_k = (1/\beta_k)(\xi^{k-1} - \xi^k)$, to calculate $\mathbf{w}_{k+1}$, formula (1.6) can be written in the equivalent form

$$\mathbf{w}_{k+1} = -\frac{1}{\beta_k}S\xi^k - \sum_{i=1}^{k} c_{k+1,i}\mathbf{w}_i, \quad c_{k+1,i} = -\frac{1}{\beta_k}\frac{(S\xi^k,\ \mathbf{w}_i)_D}{\|\mathbf{w}_i\|_D^2},$$

$$i=\overline{1,\ k}, \tag{1.8}$$

from which it immediately follows that

$$\beta_{k+1} = \frac{(\xi^k, \mathbf{w}_{k+1})_D}{\|\mathbf{w}_k\|_D^2} = -\frac{1}{\beta_k}\frac{(S\xi^k, \xi^k)_D}{\|\mathbf{w}_k\|_D^2} \neq 0.$$

In addition, it follows from

$$(S\xi^k, \mathbf{w}_i)_D = (\xi^k, S\mathbf{w}_i)_D = (\xi^k, \mathbf{w}_{i+1} + \sum_{j=1}^{i} b_{i+1,j}\mathbf{w}_j)_D$$

that $c_{k+1,i} = 0$ for $i < k$ and accordingly (1.8) has the form

$$\mathbf{w}_{k+1} = -(1/\beta_k)(S\xi^k - c_{k+1}\mathbf{w}_k),$$

$$c_{k+1} = -(1/\beta_k)\frac{(S\xi^k, \mathbf{w}_k)_D}{\|\mathbf{w}_k\|_D^2}.$$

Putting all the foregoing together and using for the D-orthogonal vectors $(-\beta_{k-1}\mathbf{w}_k)$ the same notation $\mathbf{w}_k$, $k = \overline{2, p}$, we arrive at the following two-term formulas for constructing the D-orthogonal basis for $\mathscr{K}_p(\xi^0)$:

$$\mathbf{w}_k = \begin{cases} S\xi^0, & k = 1, \\ S\xi^{k-1} - \alpha_k\mathbf{w}_{k-1}, & k > 1, \end{cases} \quad \alpha_k = \frac{(S\xi^{k-1}, \mathbf{w}_{k-1})_D}{\|\mathbf{w}_{k-1}\|_D^2}.$$
$$\xi^k = \xi^{k-1} - \beta_k\mathbf{w}_k, \quad \beta_k \frac{(\xi^{k-1}, \mathbf{w}_k)_D}{\|\mathbf{w}_k\|_D^2}, \quad k = \overline{1, p} \tag{1.9}$$

which leads to the vector $\xi^p = 0$.

From the above reasoning we readily obtain

Theorem 1.1. *For* $\mathbf{w}_k$ *and* ξ^k, $k = \overline{1, p}$, *of process* (1.9) *the following orthogonality properties are satisfied:*

$$(\mathbf{w}_k, \mathbf{w}_l)_D = 0, \quad l = \overline{1, k-1},$$

$$(\xi^k, \mathbf{w}_l)_D = 0, \quad l = \overline{1, k},$$

$$(S\xi^k, \xi^l)_D = 0, \quad l = \overline{1, k-1}.$$

Corollary. *The coefficients* α_k *and* β_k *of process* (1.9) *can be calculated from the formulas*

$$\alpha_k = -\frac{\|\xi^{k-1}\|_{DS}^2}{\|\xi^{k-2}\|_{DS}^2}, \quad \beta_k = \frac{\|\xi^{k-1}\|_{DS}^2}{\|\mathbf{w}_k\|_D^2},$$

where $\|\cdot\|_{DS}$ *is the norm generated in* U_A *by the matrix* DS.

Suppose a vector $\mathbf{u}^0 \in E_N$ is such that $\xi^0 = A\mathbf{u}^0 - \mathbf{f} \in U_A$. Then the solution $\mathbf{u}^*$ of the original system (1.1) can be found by the fol-

lowing method:

$$\mathbf{g}_k = \begin{cases} H\xi^0, & k=1, \\ H\xi^{k-1} - \alpha_k \mathbf{g}_{k-1}, & k>1, \end{cases} \quad \alpha_k = -\frac{\| \xi^{k-1} \|^2_{DS}}{\| \xi^{k-2} \|_{DS}}$$
$$\mathbf{u}^k = \mathbf{u}^{k-1} - \beta_k \mathbf{g}_k, \quad \beta_k = \frac{\| \xi^{k-1} \|^2_{DS}}{\| \mathbf{w}_k \|^2_D}, \quad k = \overline{1, p}, \tag{1.10}$$

where $\mathbf{w}_k = A\mathbf{g}_k$, $\xi^k = A\mathbf{u}^k - \mathbf{f}$ are residual vectors and $\mathbf{u}^* = \mathbf{u}^p$. Note that in the case of the singular matrix A no unique vector $\mathbf{u}^*$ can be found for given matrices H and D and a given subspace U_A, i.e. it depends on the choice of the initial vector $\mathbf{u}^0$.

If the matrix A of (1.1) is symmetric and positive definite, $H = I$ (I is the identity matrix), $D = A^{-1}$, and $U_A = E_N$, then the corresponding method (1.10), whose formulas have the form

$$\mathbf{g}_k = \begin{cases} \xi^0, & k=1, \\ \xi^{k-1} - \alpha_k \mathbf{g}_{k-1}, & k>1, \end{cases} \quad \alpha_k = -\frac{\| \xi^{k-1} \|^2}{\| \xi^{k-2} \|^2},$$
$$\mathbf{u}^k = \mathbf{u}^{k-1} - \beta_k \mathbf{g}_k, \quad \beta_k = \frac{\| \xi^{k-1} \|^2}{\| \mathbf{g}_k \|^2_A}, \quad k = \overline{1, p}, \tag{1.11}$$

is the well-known conjugate gradient method [1.1]. It is therefore natural to call method (1.10) a generalized conjugate gradient method in a subspace too.

In the case where $U_A = E_N$, $\det A \neq 0$ and, for D and S, the assumptions made above hold, i.e. D and DS are symmetric and positive definite, we have [1.6] the following

Theorem 1.2. *The generalized conjugate gradient method* (1.10) *is equivalent to the classical conjugate gradient method* (1.11) *applied to solve the system*

$$\tilde{A}\mathbf{v} = \tilde{\mathbf{f}} \tag{1.12}$$

with symmetric positive definite matrix

$$\tilde{A} = K^{-1/2} A^T D A K^{-1/2} \tag{1.13}$$

and vectors $\tilde{\mathbf{f}} = K^{-1/2} A^T D \mathbf{f}$ *and* $\mathbf{v} = K^{1/2}\mathbf{u}$, *where*

$$K = A^T D H^{-1} = A^T D (DAH)^{-1} DA \tag{1.14}$$

is a symmetric positive definite matrix.

The *proof* of the theorem is effected through a direct substitution of the expressions for the matrix $\tilde{A}$ and the vectors $\tilde{\mathbf{f}}$ and $\mathbf{v}$ into formulas (1.11).

Using Sec. 15 of [1.1] (see also [1.10] and [1.11]), we introduce more compact three-term formulas of the generalized conjugate gradient method (1.10) which do not involve calculating vectors g_k

and $\mathbf{w}_k = A\mathbf{g}_k$. Let $k < p$. Then the equations

$$\xi^{k+1} = \xi^k - \beta_k A\mathbf{g}_{k+1}, \quad \xi^k = \xi^{k-1} - \beta_k A\mathbf{g}_k,$$
$$A\mathbf{g}_{k+1} = S\xi^k - \alpha_{k+1} A\mathbf{g}_k$$

yield

$$\frac{\xi^{k+1} - \xi^k}{\beta_{k+1}} = -S\xi^k - \alpha_{k+1}\frac{\xi^k - \xi^{k-1}}{\beta_k}$$

or equivalently

$$\xi^{k+1} = \xi^k - \beta_{k+1} S\xi^k - (\alpha_{k+1}\beta_{k+1}/\beta_k)(\xi^k - \xi^{k-1}).$$

Hence

$$\mathbf{u}^k = \mathbf{u}^{k-1} - (1/q_k)[H\xi^{k-1} - e_k(\mathbf{u}^{k-1} - \mathbf{u}^{k-2})], \tag{1.15}$$

where

$$q_k = \frac{1}{\beta_k}, \quad e_{k-1} = \begin{cases} 0, & k = 1, \\ -\alpha_k/\beta_{k-1}, & k > 1. \end{cases}$$

We show that the quantities q_k and e_{k-1} can be evaluated from the following formulas ($e_0 = 0$):

$$q_k = \frac{\| S\xi^{k-1} \|_D^2}{\| \xi^{k-1} \|_{DS}^2} - e_{k-1}, \quad e_k = q_k \frac{\| \xi^k \|_{DS}^2}{\| \xi^{k-1} \|_{DS}^2}, \quad k = \overline{1, p}. \tag{1.16}$$

Indeed, using the properties of $A\mathbf{g}_k$ and ξ^k we get

$$\begin{aligned} \| A\mathbf{g}_k \|_D^2 &= (S\xi^{k-1} - \alpha_k A\mathbf{g}_{k-1}, \quad A\mathbf{g}_k)_D = (S\xi^{k-1}, \quad A\mathbf{g}_k)_D \\ &= (S\xi^{k-1}, \quad S\xi^{k-1} - \alpha_k A\mathbf{g}_k)_D = \| S\xi^{k-1} \|_D^2 \\ &\quad - (\alpha_k/\beta_{k-1})(S\xi^{k-1}, \quad \xi^{k-2} - \xi^{k-1}) \\ &= \| S\xi^{k-1} \|_D^2 - e_{k-1} \| \xi^{k-1} \|_{DS}^2 \end{aligned}$$

from which we at once obtain formulas (1.16):

$$q_k \quad \frac{\| A\mathbf{g}_k \|_D^2}{\| \xi^{k-1} \|_{DS}^2} = \frac{\| S\xi^{k-1} \|_D^2}{\| \xi^{k-1} \|_{DS}^2} - e_{k-1},$$

$$e_k = -\alpha_{k+1}/\beta_k = -\alpha_{k+1} q_k = q_k \frac{\| \xi^k \|_{DS}^2}{\| \xi^{k-1} \|_{DS}^2}.$$

Now we consider briefly the estimates of the rate of convergence of the method under consideration.

Denote by $\mathbf{u}^*$ the solution of system (1.1) which we obtain for a given $\mathbf{u}^0$ in $p = p(\xi^0)$ steps of the generalized conjugate gradient method in U_A and consider a sequence of error vectors $\mathbf{z}^k = \mathbf{u}^k - \mathbf{u}^*$. It is not difficult to show that $\mathbf{z}^k$, $k = \overline{0, p}$, are in the subspace HU_A which is the span of the A^TDA-orthonormal system of

eigenvectors φ_i, $i = \overline{1, t}$, of the matrix HA that correspond to its positive eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_t$. These eigenvectors are related to the corresponding previously determined eigenvectors ψ_i of the matrix $S = AH$ by the equations $\psi_i = A\varphi_i$, $i = \overline{1, t}$.

Suppose that the eigenvalues λ_i, $i = \overline{1, t}$, lie in the interval $[a, b]$, where $a < b$ are some positive integers, and denote by $T_k(x)$ Chebyshev polynomials of the first kind of order k ($k \geqslant 0$):

$$T_k(x) = \begin{cases} \cos(k \cos^{-1} x), & |x| \leqslant 1, \\ \cosh(k \cosh^{-1} x), & |x| > 1, \end{cases}$$

or equivalently

$$T_k(x) = (1/2)\left[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k\right].$$

Then we have ([1.4], [1.12], [1.13])

Theorem 1.3. *Under the assumptions made, for the sequence* $D_1 = D^T A D$*-norms of the error vectors* $\mathbf{z}^k$ *of the generalized conjugate gradient method* (1.10) *we have the estimate*

$$\| \mathbf{z}^k \|_{D_1} \leqslant \frac{1}{T_k[(b+a)/(b-a)]} \cdot \| z^0 \|_{D_1}, \tag{1.17}$$

where

$$\begin{aligned} &T_k[(b + a)/(b - a)] = (1 + q^{2k})/2q^k, \\ &q = (1 - \nu^{1/2})/(1 + \nu^{1/2}), \quad \nu = a/b. \end{aligned} \tag{1.18}$$

Corollary 1. *To minimize the* D_1*-norm of the vector of initial error by a factor of* $1/\varepsilon$ ($\varepsilon < 1$) *it suffices to carry out* k_ε *steps of method* (1.10), *where* k_ε *is the minimum whole number satisfying the inequality*

$$k \geqslant \frac{\ln[1/\varepsilon + \sqrt{1/\varepsilon - 1}]}{\ln[(1+\nu^{1/2})/(1-\nu^{1/2})]}. \tag{1.19}$$

Corollary 1.2. *If* ν, $\varepsilon \ll 1$, *then* k_ε *is evaluated by the following approximate formula*:

$$k_\varepsilon \cong (\nu^{-1/2}/2) \ln(2/\varepsilon). \tag{1.20}$$

1.2. SIMPLEST APPLICATIONS

In this section we consider some simplest applications of the general theory of the generalized conjugate gradient method in a subspace when for given matrices A and H the subspace U_A is defined by the relation

$$U_A = \operatorname{Im}[(I - AH)A] \subseteq \operatorname{Im}(I - AH) \cap \operatorname{Im} A. \tag{1.21}$$

Then the initial vector of method (1.10) can be chosen by using one iteration of the corresponding stationary iterative method, i.e. by

the formula

$$\mathbf{u}^0 = \mathbf{v} - H(A\mathbf{v} - \mathbf{f}) \tag{1.22}$$

for some, in general arbitrary, vector $\mathbf{v} \in E_N$. In this case the vector

$$\xi^0 = A\mathbf{u}^0 - \mathbf{f} = (I - AH)(A\mathbf{v} - \mathbf{f})$$

will belong to the chosen subspace U_A.

Let A and H be symmetric positive semidefinite matrices satisfying the requirement $\ker H \cap \operatorname{Im} A = 0$. Then, since $U_A \subseteq \operatorname{Im} A$ and, for any $\xi, \eta \in U_A$,

$$(DAH\,\xi, \eta) = (H\xi, \eta),$$

the matrix $D = A^+$ and the corresponding matrix DAH are self-adjoint positive definite in U_A. Thus, to solve (1.1) one can apply (1.10), with given D, A, and H, whose two-term implementation formulas have the form

$$\begin{gathered} g_k = \begin{cases} H\xi^0, & k = 1, \\ H\xi^{k-1} - \alpha_k g_{k-1}, & k > 1, \end{cases} \quad \alpha_k = -\frac{\|\xi^{k-1}\|_H^2}{\|\xi^{k-2}\|_H^2}, \\ \mathbf{u}^k = \mathbf{u}^{k-1} - \beta_k g_k, \quad \beta_k = \|\xi^{k-1}\|_H^2 / \|g_k\|_A^2. \end{gathered} \tag{1.23}$$

Similarly, q_k and e_k of (1.15) will be evaluated from

$$q_k = \frac{\|H\xi^{k-1}\|_A^2}{\|\xi^{k-1}\|_H^2} - e_{k-1}, \quad e_k = q_k \frac{\|\xi^k\|_H^2}{\|\xi^{k-1}\|_H^2}. \tag{1.24}$$

It is obvious that when $H = I$ we arrive at the usual conjugate gradient method (1.11).

Now let A be an arbitrary matrix, H_1 and H_2 be symmetric positive semidefinite matrices satisfying the requirements $\ker H_2 \cap \operatorname{Im} A = 0$ and $\ker H_1 \cap \operatorname{Im} A^T = 0$, $H = H_1 A^T H_2$ and $U_A = \operatorname{Im} A$. Then, since for any $\xi, \eta \in \operatorname{Im} A$

$$(DAH\,\xi, \eta) = (H_2\xi, \eta),$$

it is not hard to see that matrices $D = AH_1A^{T+}$ and DAH are self-adjoint positive definite in U_A. After simple transformations it can be shown that the implementation formulas for the corresponbing method (1.10) are of the form

$$y_k = \begin{cases} A^T H_2 \xi^0, & k = 1, \\ A^T H_2 \xi^{k-1} - \alpha_k y_{k-1}, & k > 1, \end{cases} \quad \alpha_k = -\frac{\|\xi^{k-1}\|_{H_2}^2}{\|\xi^{k-2}\|_{H_2}^2}, \tag{1.25}$$

$$g_k = H_1 y_k, \quad \mathbf{u}^k = \mathbf{u}^{k-1} - \beta_k g_k, \quad \beta_k = \frac{\|\xi^{k-1}\|_{H_2}^2}{\|y_k\|_{H_1}^2}.$$

Similarly, q_k and e_k of (1.15) are evaluated from

$$q_k = \frac{\| A^T H_2 \xi^{k-1} \|^2_{H_1}}{\| \xi^{k-1} \|^2_{H_2}} - e_{k-1}, \quad e_k = q_k \frac{\| \xi^k \|^2_{H_2}}{\| \xi^{k-1} \|^2_{H_2}}. \tag{1.26}$$

Now consider the case where A is a symmetric positive semidefinite block matrix of the following form:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{1.27}$$

where A_{11} and A_{22} are positive definite matrices of order N_1 and N_2, respectively ($N = N_1 + N_2$). We choose $H = B^{-1}$, where

$$B = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \tag{1.28}$$

and define the matrix $C = B - A$. Then $U_A \subseteq \operatorname{Im} C \cap \operatorname{Im} A$ and assuming $D = A^+$ it is not hard to see that for any $\xi, \eta \in U_A$

$$(AH\xi, \eta)_D = (H\xi, \eta) = (\xi, H\eta),$$

i.e. the matrix DAH is self-adjoint in U_A. Moreover, if we represent $\xi \in U_A$ as $\xi = \begin{bmatrix} 0 \\ \xi_2 \end{bmatrix}$, where $\xi_2 \in E_{N_2}$, then for any $\xi \neq 0$ ($\xi_2 \neq 0$) we have

$$(DAH\,\xi, \xi) = (H\xi, \xi) = (A_{22}^{-1}\xi_2, \xi_2) > 0$$

and hence the matrix DAH is positive definite in U_A (see [1.14]). Thus, to solve (1.1) one can apply the generalized conjugate gradient method (1.10) with given H and D in a chosen subspace U_A. We write out the implementation formulas for this method, we shall yet need them in Sec. 1.5.

Suppose that for some $k \geqslant 0$ we know vectors $\psi^k = C\mathbf{u}^k$ and $\xi^k = A\mathbf{u}^k - \mathbf{f}$ or, to be more precise, vectors $\psi_2^k, \xi_2^k \in E_{N_2}$ from the presentation $\psi^k = \begin{bmatrix} 0 \\ \psi_2^k \end{bmatrix}$, $\xi^k = \begin{bmatrix} 0 \\ \xi_2^k \end{bmatrix}$. Then the vector $\mathbf{u}^k$ can be reconstructed by solving the equation

$$B\mathbf{u}^k = C\mathbf{u}^k + \xi^k + \mathbf{f} \tag{1.29}$$

or equivalently by solving the system

$$\begin{aligned} A_{11}\mathbf{u}_1^k + A_{12}\mathbf{u}_2^k &= \mathbf{f}_1, \\ A_{22}\mathbf{u}_2^k &= \mathbf{f}_2 + \psi_2^k + \xi_2^k, \end{aligned} \tag{1.30}$$

although it would be more correct to evaluate $\mathbf{v}^k$, which is the solution of the system

$$B\mathbf{v}^k = \mathbf{f} + \xi^k, \tag{1.31}$$

instead of $\mathbf{u}^k$, since in this case we have

$$\| \mathbf{v}^k - \mathbf{u}^* \|_A < \| \mathbf{u}^k - \mathbf{u}^* \|_A \tag{1.32}$$

in the space of error vectors U.

Thus in the course of the implementation of method (1.10) it is sufficient for us to evaluate only the components ψ_2^k and ξ_2^k of ψ^k and ξ^k and after $k = \hat{k}$ steps (for example, $\hat{k} = k_\varepsilon$ or $\hat{k} = p$) either to use (1.30) to reconstruct the entire vector $\mathbf{u}^k$ or to use (1.31) to find a more accurate approximation $\mathbf{v}^k$ to the solution $\mathbf{u}^*$ of the original system (for $\hat{k} = p$ we have $\mathbf{u}^{\hat{k}} = \mathbf{v}^{\hat{k}} = \mathbf{u}^*$). The corresponding implementation formulas for method (1.10) have the form ($\nu = 0$):

$$\begin{aligned}
&C\mathbf{u}^0 = CB^{-1}\mathbf{f}, \quad \xi^0 = -C\mathbf{u}^0,\\
&\varphi^k = A_{21}A_{11}^{-1}A_{12}A_{22}^{-1}\xi_2^{k-1},\\
&\mathbf{y}_2^k = \begin{cases} \varphi^1, & k = 2,\\ \varphi^k - \alpha_k \mathbf{y}_2^{k-1}, & k > 1, \end{cases}\\
&\mathbf{w}_2^k = \begin{cases} \xi_2^0 - \varphi^1, & k = 1,\\ \xi_2^{k-1} - \varphi^k - \alpha_k \mathbf{w}_2^{k-1}, & k > 1, \end{cases}\\
&\alpha_k = -\left\{ \frac{\| \xi_2^{k-1} \|^2_{A_{22}^{-1}}}{\| \xi_2^{k-2} \|^2_{A_{22}^{-1}}} \right.\\
&\psi_2^k = \psi_2^{k-1} - \beta_k \mathbf{y}_2^k, \quad \xi_2^k = \xi_2^{k-1} - \beta_k \mathbf{w}_2^k,\\
&\beta_k = \| \xi_2^{k-1} \|^2_{A_{22}^{-1}} / (\mathbf{w}_2^k, A_{22}^{-1}\xi_2^{k-1}),
\end{aligned} \tag{1.33}$$

where $\mathbf{y}_2^k, \mathbf{w}_2^k \in E_{N_2}$ are the components of vectors $\mathbf{y}^k = C\mathbf{g}_k = \begin{bmatrix} 0 \\ \mathbf{y}_2^k \end{bmatrix}$ and $\mathbf{w}^k = A\mathbf{g}_k = \begin{bmatrix} 0 \\ \mathbf{w}_2^k \end{bmatrix}$ from $U_A \subseteq \operatorname{Im} C$. It is obvious that to implement (1.33) it suffices to store (ignoring the procedure for finding the vector φ^k from a given vector ξ_2^{k-1}) five vectors of dimension N_2.

The two-term formulas of method (1.15) similar to (1.33) can be written as

$$\begin{aligned}
&\varphi^k = A_{21}A_{11}^{-1}A_{12}A_{22}^{-1}\xi_2^{k-1},\\
&\psi_2^k = \psi_2^{k-1} - (1/q_k)\,[\varphi^k - e_{k-1}(\psi_2^{k-1} - \psi_2^{k-2})],\\
&\xi_2^k = \xi_2^{k-1} - (1/q_k)\,[\xi_2^k - \varphi_2^k - e_{k-1}\,(\xi_2^{k-1} - \xi_2^{k-2})],
\end{aligned} \tag{1.34}$$

where $e_0 = 0$,

$$q_k = \frac{(A_{22}^{-1}\xi_2^{k-1},\ \xi_2^{k-1} - \varphi^k)}{\| \xi_2^{k-1} \|^2_{A_{22}^{-1}}} - e_{k-1},$$

$$e_k = q_k \frac{\| \xi_2^k \|^2_{A_{22}^{-1}}}{\| \xi_2^{k-1} \|_{A_{22}^{-1}}}, \tag{1:35}$$

To implement (1.34), (1.35) it is also sufficient to store simultaneously only five vectors of dimension N_2.

Thus, the most important procedure in terms of the amount of computational work required for the implementation of each step in methods (1.33) and (1.34), (1.35) is (especially in the case $N_2 \ll N$) to calculate φ^k from a given vector ξ_2^{k-1}. We shall return to this question in Sec. 1.5.

It is clear that methods (1.33) and (1.34), (1.35) can be interpreted in another way. We can treat them as the generalized conjugate gradient method applied to the solution of a transformed, implicitly given system

$$\tilde{A}\mathbf{v} = \mathbf{F}, \tag{1.36}$$

where $\tilde{A} = A_{22} - A_{21}A_{11}^{-1}A_{12}$, $\mathbf{v} = \mathbf{u}_2$ and $\mathbf{F} = \mathbf{f}_2 - A_{21}A_{11}^{-1}\mathbf{f}_1$. Chosen as H is $\tilde{H} = A_{22}^{-1}$ and the analogue of U_A is $\tilde{U}_A = \operatorname{Im}[(I - \tilde{A}\tilde{H})\tilde{A}] \subseteq E_{N_2}$, the implementation formulas for (1.33), for example, assuming the simpler form (1.23).

Thus, in the situation we have considered the methods in a subspace can be given a different, more convenient and simpler interpretation using the idea of the methods of [1.15]. But in our opinion based on experience such an interpretation is not always sufficiently obvious or visual. In addition, the concept presented in Sec. 1.1 seems to our minds to be more general, from the viewpoint of uniting iterative methods different in their properties into a single class, as well as more constructive when developing new methods aimed at particular applied problems.

We conclude this section by discussing briefly the estimation of the rate of convergence of method (1.33). It is known that the eigenvalues of the transition matrix $T = I - B^{-1}A$ (where I is the identity matrix and the matrices $B = H^{-1}$ and A have been defined above) of the block Gauss-Seidel method lie in the interval $[0, \rho(T)]$, where $\rho(T)$ is the spectral radius of the matrix T. This matrix has a complete system of eigenvectors and $\rho(T) \leqslant 1$, with $\rho(T) = 1$ only if $\det A = 0$. It at once follows that U_A is the span of the eigenvectors of AH corresponding to its eigenvalues not equal to zero or unity. To put it another way, for the method under consideration it

is possible to set

$$a = \min_{\substack{\lambda \in \sigma(HA) \\ \lambda \neq 0}} \lambda, \quad b = \max_{\substack{\lambda \in \sigma(HA) \\ \lambda \neq 1}} \lambda, \tag{1.37}$$

where $\sigma(HA)$ is the set of all eigenvalues of HA, with $b < 1$. It follows, in particular, that when $a \ll 1$ and $|1 - b| \ll 1$ instead of (1.20) one can use for k_ε the approximate formula

$$k_\varepsilon \cong (a^{-1/2}/2) \ln(2/\varepsilon), \tag{1.38}$$

where $a = 1 - \rho(T)$ (if $\det A \neq 0$).

1.3. THE METHOD OF A SIMPLE ITERATION FOR THE KINETIC TRANSPORT EQUATION

Given a domain $G \times \Omega$, where G is a unit cube and Ω is a unit sphere, consider the one-velocity integrodifferential kinetic transport equation [1.16]

$$(\mathbf{s}, \operatorname{grad}) \mathbf{u} + \sigma \mathbf{u} = \frac{\sigma_t}{4\pi} \int_\Omega \mathbf{u}\, d\Omega' + \mathbf{f} \tag{1.39}$$

with the homogeneous boundary condition

$$\mathbf{u} = 0 \text{ for } \mathbf{x} = (x_1, x_2, x_3) \in \partial G, \quad (\mathbf{s}, \boldsymbol{\nu}) < 0. \tag{1.40}$$

Here $\sigma = \sigma(\mathbf{x})$ and $\sigma_t = \sigma_t(\mathbf{x})$ are nonnegative continuous functions satisfying the inequality $\sigma \geqslant \sigma_t$, $f = f(\mathbf{x}, \mathbf{s})$ is a continuous function and $\boldsymbol{\nu}$ is the vector of the outward normal to the boundary ∂G of G.

We construct a uniform square grid with a mesh-size $h = 1/(n+1)$ in G and a grid with nodes $\mathbf{s}_l$, $l = \overline{1, 2m}$, where m is some integer, satisfying the condition $\mathbf{s}_l = -\mathbf{s}_{m+l}$, $l = \overline{1, m}$, on the unit sphere Ω. We then approximate the original problem (1.39), (1.40) using the simplest system of grid equations [1.17] which in matrix notation has the form

$$L\mathbf{u} = S\mathbf{u} + \mathbf{f}, \tag{1.41}$$

where the matrix L corresponds to the differential operator on the left of (1.39) and S corresponds to the integral operator at the right. With an appropriate ordering of the unknowns and equations the matrices L and S can be represented as

$$L = \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda^T \end{bmatrix} + \Sigma, \quad S = \Sigma_t \begin{bmatrix} S_1 & S_1 \\ S_1 & S_1 \end{bmatrix}. \tag{1.42}$$

Here Λ is a block diagonal lower triangular matrix, $\Sigma = \tilde{\Sigma} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_t = \tilde{\Sigma}_t \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, where $\tilde{\Sigma}$ and $\tilde{\Sigma}_t$ are diagonal matrices with nonnegative diagonal elements, and S_1 is a symmetric positive semidefinite matrix of rank n^3. The above matrices have the following properties:

(i) $S_1 = S_1^2/2$, i.e. the matrix $\sqrt{2}S_1$ is an orthogonal projector;
(ii) the matrices L and $L - S$ are positive definite,
(iii) L and $L - S$ are M-matrices;
(iv) the splitting $A = L - S$ is regular [1.18] and consequently $\rho(L^{-1}S) < 1$.

One of the most widely used methods of solving (1.41) is the so-called method of a simple iteration

$$L\mathbf{u}^k = S\mathbf{u}^{k-1} + \mathbf{f}, \quad k = 1, 2, \ldots \tag{1.43}$$

every step of which requires $O(N)$ arithmetic operations, where $N = 2mn^3$. Obviously, for all $k \geqslant 1$

$$\xi^k = A\mathbf{u}^k - \mathbf{f} = S\mathbf{u}^{k-1} - S\mathbf{u}^k \in \operatorname{Im} S.$$

We choose $U_A = \operatorname{Im}(I - L^{-1}S) = \operatorname{Im} S$. Then the following lemma is valid.

Lemma 1.2. *Matrices A^{-1}, L^{-1}, and S^+ are self-adjoint and positive definite operators in U_A.*

Proof. Indeed each of the matrices can be represented in the block form

$$\begin{bmatrix} D_1 & D_2 \\ D_2 & D_1^T \end{bmatrix}$$

where D_1 and D_2 are some matrices of order $N/2$, with D_2 being symmetric. But for any ξ $\eta \in U_A$ we then have

$$(D\xi, \eta) = (\hat{D}\xi, \eta),$$

where $\hat{D} = \frac{1}{4}(D_1 + 2D_2 + D_1^T) \otimes \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is a symmetric positive definite matrix in U_A.

Lemma 1.3. *Let D be one of the matrices A^{-1}, L^{-1}, or S^+. Then the matrix AL^{-1} is D-self-adjoint and D-positive definite in U_A.*

The *proof* of the lemma makes use of the structure of U_A and of simple matrix transformations for every particular choice of D.

A direct consequence of the last two lemmas is

Theorem 1.4. *To solve* (1.41) *we use the generalized conjugate gradient method in $U_A = \operatorname{Im} S$ with $H = L^{-1}$ and D equal to A^{-1}, L^{-1}, or S^+.*

We write out computational formulas of the method for $D = A^{-1}$. Similarly to the last method of the preceding section we set the following problem: given $k = \hat{k}$ find first $S\mathbf{u}^{\hat{k}}$, $\xi^{\hat{k}} \in U_A$, and only after that find the required vector $\mathbf{u}^{\hat{k}}$ or vector $\mathbf{v}^{\hat{k}}$ as a solution of the system

$$L\mathbf{v}^{\hat{k}} = S\mathbf{u}^{\hat{k}} + \mathbf{f},$$

satisfying, by virtue of $\rho(L^{-1}S) < 1$,

$$\| A\mathbf{v}^{\hat{k}} - \mathbf{f} \|_D < \| A\mathbf{u}^{\hat{k}} - \mathbf{f} \|_D.$$

Thus for $k = 1, 2, \ldots, \hat{k}$ we have

$$\begin{aligned} S\mathbf{u}^k &= S\mathbf{u}^{k-1} - (1/q_k)\,[SL^{-1}\xi^{k-1} - e_{k-1}(S\mathbf{u}^{k-1} - S\mathbf{u}^{k-2})], \\ \xi^k &= \xi^{k-1} - (1/q_k)\,[\xi^{k-1} - SL^{-1}\xi^{k-1} - e_{k-1}(\xi^{k-1} - \xi^{k-2})], \end{aligned} \tag{1.44}$$

where $e_0 = 0$ and the other values of e_k and q_k are found using the following formulas:

$$\begin{aligned} q_k &= \frac{(SL^{-1}\xi^{k-1},\ (I - SL^{-1})\xi^{k-1})_{\Sigma_t^+}}{(\xi^{k-1},\ SL^{-1}\xi^{k-1})_{\Sigma_t^+}} - e_{k-1} \\ e_k &= q_k\,(SL^{-1}\xi^k,\ \xi^k)_{\mathbf{g}_t^+}/(SL^{-1}\xi^{k-1},\ \xi^{k-1})_{\Sigma_t^+}. \end{aligned} \tag{1.45}$$

In deriving the last formulas we have used the fact that $S^+S = \Sigma_t^+ S$ and therefore for any two vectors $\xi \in U_A$ and $\eta \in E_N$

$$(\xi,\ \eta) = (\xi,\ S\eta)_{\Sigma_t^+}.$$

Some words about the computational characteristics of method (1.44), (1.45). Since the rank of S is at most n^3, to implement the method it suffices to store five of n^3-dimensional vectors and use at each step only $O(n^3)$ arithmetic operations. At the same time to implement the procedure $\xi \to SL^{-1}\xi$ requires $O(N)$ arithmetic operations. Thus the use of formulas (1.44), (1.45) of the conjugate gradient method to optimize the simple iteration method (1.43), while increasing the required storage by two and a half times, does not lead to any considerable increase in the number of arithmetic operations. The rate of convergence of method (1.44), (1.45) is naturally much higher (especially if $\rho(L^{-1}S)$ is close to unity) than that of the simple iteration method (1.43).

Extensions of the method presented to other cases, in particular to problems with nonisotropic dispersion, are discussed in [1.17], [1.19], and [1.20].

1.4. FICTITIOUS COMPONENTS METHOD

Consider a system of linear algebraic equations

$$A_1\mathbf{u}_1 = \mathbf{f}_1 \tag{1.46}$$

with symmetric positive semidefinite matrix A_1 of order N_1 and vector $\mathbf{f}_1 \in \operatorname{Im} A_1$ and suppose that its solution uses the iterative method

$$B_1(\mathbf{u}_1^k - \mathbf{u}_1^{k-1}) = -\tau_k(A_1\mathbf{u}_1^{k-1} - \mathbf{f}_1) \tag{1.47}$$

with some symmetric positive definite matrix B_1 and a parameter sequence τ_k. Then, to make easier the choice of readily invertible matrices B, in particular matrices of special structure (this is especially important for systems of grid equations), it was suggested in [1.21] that system (1.46) should first be transformed to system (1.1) with vector $\mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ 0 \end{bmatrix} \in E_N$ and matrix

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \tag{1.48}$$

of order $N > N_1$, where A_2 is some, in general, arbitrary symmetric positive semidefinite matrix of order $N_2 = N - N_1$, and then one should apply the iterative method

$$B(\mathbf{u}^k - \mathbf{u}^{k-1}) = -\tau_k(A\mathbf{u}^k - \mathbf{f}) \tag{1.49}$$

with a chosen symmetric and positive definite matrix B to the resulting system $A\mathbf{u} = \mathbf{f}$.

The procedure described is known as the method of fictitious or false components. In [1.22] it was pointed out that other variants of the fictitious components method are possible, for example, one with the matrix

$$A = \begin{bmatrix} A_1 & 0 \\ A_{12} & A_2 \end{bmatrix}. \tag{1.50}$$

Specific applications of the fictitious components method with transformation (1.48) to the solution of elliptic equations with natural boundary conditions were studied by Astrakhantsev ([1.23], [1.24]) and with principal boundary conditions by Matsokin [1.25]. The fictitious components method with transformations of the type (1.50) for elliptic equations with principal boundary conditions (close relations with the capacitance matrix method [1.26], [1.27] are meant here) is dealt with in [1.28] and [1.29]. Close links between approaches (1.48) and (1.50) were noted by Matsokin [1.30].

In [1.21] and [1.22] the following theorem was proved.

Theorem 1.5. *Suppose that for the matrix A from* (1.48) *and a matrix B the above assumptions hold,*

$$a = \min_{\substack{\lambda \in \sigma(B^{-1}A) \\ \lambda \neq 0}} \lambda, \quad b = \max_{\lambda \in \sigma(B^{-1}A)} \lambda, \tag{1.51}$$

and let A_1 and B be fixed. Then the maximum ratio a/b is reached at $A_2 = 0$.

Thus in terms of the maximum ratio a/b, which determines the rate of convergence of Chebyshev iterative methods and that of the generalized conjugate gradient method in a subspace (the closer a/b is to unity, the higher the rate of convergence is), the optimum variant of the fictitious components method is $A_2 = 0$. If therefore U_A is given by formula (1.21), the last N_2 components of vectors $\xi \in U_A$ will be zero and the form of the first N_1 components will be determined by the structure of the first N_1 rows of the matrix $B - A$. We consider how these remarks are used in a particular situation.

Suppose in a two-dimensional bounded domain Ω with piecewise smooth boundary $\partial\Omega$ the following equation is given

$$-\Delta u + cu = f, \tag{1.52}$$

where $c = \text{const} > 0$, with the boundary conditions

$$\alpha \frac{\partial u}{\partial \nu} + \beta u = g \quad \text{on } \partial\Omega, \tag{1.53}$$

where ν is the vector of the outward normal to $\partial\Omega$, α is a given positive function and β is a given nonnegative function.

We enclose the domain Ω in some open rectangle G with sides parallel to the coordinate axes and consider the problem

$$\begin{aligned} -\Delta v + cv &= F \text{ in } G, \\ \partial v/\partial \nu &= 0 \text{ on } \partial G, \end{aligned} \tag{1.54}$$

where ν is the vector of the outward normal to the boundary ∂G of G.

We then construct in G a uniform rectangular grid, triangulate it in the standard way and assume that $\partial\Omega$ belongs to the union of sides of elementary triangles (see Fig. 1.1).

Under the assumptions made, upon discretizing problems (1.52), (1.53), and (1.54) we accordingly arrive at system (1.46) and obtain the matrix

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{1.55}$$

where B_{ij} are $N_i \times N_j$ blocks, with $i, j = 1, 2$. Following Theorem 1.5 we get

$$A = \begin{bmatrix} A_1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{1.56}$$

It then follows from [1.23], [1.24] that under the assumptions made the ratio a/b of the quantities in (1.51) is bounded below by a constant independent of the mesh spacing, so we shall agree to consider this ratio a constant.

It is not hard to see that under the assumptions made the rows of $B_{11} - A_1$ and B_{12}, which correspond to the grid nodes in Ω, are

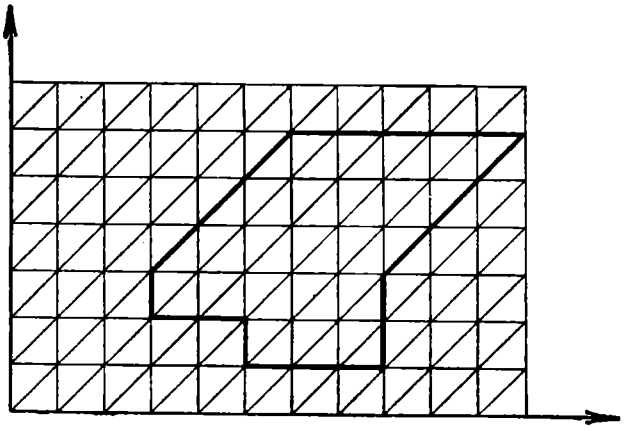


Fig. 1.1.

all zero, i.e. the nonzero rows may correspond only to the grid nodes on $\partial\Omega$.

It follows from the foregoing that only those components of any vector $\xi \in U_A$ may be nonzero which correspond to nodes on $\partial\Omega$. The total number of such nodes is $O(N^{1/2})$. Notice that the number of nonzero columns in $B_{11} - A_1$ and B_{12} is also at most $O(N^{1/2})$ and, for the matrix $B_{11} - A_1$, they may correspond only to the nodes on $\partial\Omega$ and for B_{12} only to the nodes at the boundary $\partial\Omega$ that are not in Ω.

We introduce a diagonal matrix R whose diagonal elements equal unity if given rows correspond to the nodal points on $\partial\Omega$ and are zero otherwise. Then, according to the preceding reasoning, the conjugate gradient method with matrices $H = B^{-1}$ and $D = A^*$ in a chosen subspace U_A can be implemented by the formulas

$$\begin{aligned} R\mathbf{u}^k &= R\mathbf{u}^{k-1} - (1/q_k)\,[RB^{-1}\xi^{k-1} - e_{k-1}(R\mathbf{u}^{k-1} - R\mathbf{u}^{k-2})], \\ \xi^k &= \xi^{k-1} - (1/q_k)\,[AB^{-1}\xi^{k-1} - e_{k-1}(\xi^{k-1} - \xi^{k-2})], \end{aligned} \tag{1.57}$$

where $e_0 = 0$,

$$q_k = \frac{(RB^{-1}\xi^{k-1},\ AB^{-1}\xi^{k-1})}{(RB^{-1}\xi^{k-1},\ \xi^{k-1})} - e_{k-1}, \quad e_k = q_k \frac{(RB^{-1}\xi^k,\ \xi^k)}{(RB^{-1}\xi^{k-1},\ \xi^{k-1})}. \tag{1.58}$$

It is obvious that the main computational problem at the kth step of the method is (all the rest requiring only $O(N^{1/2})$ arithmetic operations) to find from a given vector $\xi^{k-1} \in U_A$ the vectors

$BB^{-1}\xi^{k-1}$ and $AB^{-1}\xi^{k-1}$ or, to be more precise, those components of these vectors which can be nonzero. Taking into account the structure of R and the fact that

$$AB^{-1}\xi^{k-1} = \xi^{k-1} - \begin{bmatrix} B_{11} - A_1, & B_{12} \\ 0 & 0 \end{bmatrix} B^{-1}\xi^{k-1}, \tag{1.59}$$

to solve the problem set it suffices to find only those components of the vector φ^k, the solution of the system

$$B\varphi^k = \xi^{k-1}, \tag{1.60}$$

which correspond to the grid nodes on $\partial\Omega$ and the nodes at $\partial\Omega$ in $G \setminus \bar{\Omega}$. It is known ([1.6], [1.31]) that this problem can be solved by separation of variables in $O(N)$ arithmetic operations, with only $O(N^{1/2})$ quantities stored at the same time. Thus each step of method (1.57), (1.58) can be implemented in $O(N)$ arithmetic operations, while storing five data arrays whose dimension does not exceed the number of nodes on $\partial\Omega$.

The foregoing is summarized by

Theorem 1.6. *System* (1.46) *resulting from discretizing problem* (1.52), (1.53) *is solved by method* (1.48), (1.57), (1.58) *to an accuracy of* ε ($\varepsilon < 1$) *in* $O(\ln(1/\varepsilon))$ *steps to implement which it suffices to perform* $O(N \ln(1/\varepsilon))$ *arithmetic operations and to store only a finite number of arrays of dimension* $O(N^{1/2})$.

Notice that to find the initial vectors $R\mathbf{u}^0$ and ξ^0 and to reconstruct $\mathbf{u}^{\hat{k}}$ will require $O(N \ln \ln N)$ arithmetic operations if the algorithms of [1.31] and [1.32] are used.

1.5. DOMAIN DECOMPOSITION METHODS

An important class of iterative methods in subspaces comprises methods based on the idea of domain decomposition. In recent years a large number of efficient algorithms which could be optimized using the generalized conjugate gradient method [1.15], [1.29], [1.34]-[1.37] have been proposed along these lines. We shall discuss only one of them here that proposed by one of the authors in [1.34] and developed in subsequent work [1.15], [1.35]. Moreover we shall restrict ourselves to the simplest model case which allows the most economical implementation.

Let a two-dimensional domain $\bar{\Omega}$ be the union of a finite number of rectangles Ω_i, $i = \overline{1, r}$, whose sides are parallel to the coordinate axes, with different open rectangles not intersecting and the intersections of different closed rectangles parallel to only one (fixed) coordinate axis. We also assume that the region allows a uniform grid to be constructed using an appropriate variable (for example, r_1).

Consider in Ω the diffusion equation

$$-\frac{\partial}{\partial x_1}\left(P\frac{\partial u}{\partial x_1}\right)-\frac{\partial}{\partial x_2}\left(Q\frac{\partial u}{\partial x_2}\right)+Ku=f \tag{1.61}$$

with constant nonnegative values of the coefficients P, Q, and K in each rectangular subdomain Ω_i subject to the boundary condition

$$\alpha\partial u/\partial\nu+\beta u=g, \tag{1.62}$$

where f, g, α, β are given functions and ν is the direction of the outward normal to $\partial\Omega$. As to α and β it is further assumed that they are nonnegative, that the function $\alpha+\beta$ is positive on $\partial\Omega$, and that on each side of the rectangles making up Ω that are parallel to the x_1-axis either α or β is identically zero.

We construct in Ω a rectangular grid with a uniform spacing in the direction of the variable x_2 and using the simplest difference method (or variational difference method) carry out discretization (or quantization) of the original problem. As a result we arrive at the system of linear algebraic equations (1.1) with block symmetric matrix

$$A=\begin{bmatrix} \tilde{A}_{11} & & 0 & \tilde{A}_{1,\,r+1} \\ & \ddots & & \vdots \\ 0 & & \tilde{A}_{r,\,r} & \tilde{A}_{r,\,r+1} \\ \tilde{A}_{r+1,\,1} & \cdots & \tilde{A}_{r+1,\,r} & \tilde{A}_{r+1,\,r+1} \end{bmatrix}. \tag{1.63}$$

As to the blocks $\tilde{A}_{ii}$ we assume that they correspond to the similar difference approximations of the problems

$$\begin{aligned} -P_i\partial^2u/\partial x_1^2-Q_i\partial^2u/\partial x_2^2+K_iu&=f && \text{in} && \Omega_i \\ \alpha\partial u/\partial\nu+\beta u&=g && \text{on} && \partial\Omega_i\cap\partial\Omega, \\ u&=0 && \text{on} && \partial\Omega_i\cap\partial\Omega, \end{aligned} \tag{1.64}$$

where $\partial\Omega_i$ are the boundaries of subdomain Ω_i, $i=\overline{1,\,r}$. It is obvious that the matrix $\tilde{A}_{r+1,r+1}$ corresponds to the grid equations on the sides of Ω_i parallel to the x_2-axis and it has order $O(N^{1/2})$, where N is the order of A (it is assumed that the number of nodes with respect to each of the variables is $O(N^{1/2})$). It is obvious that $\tilde{A}_{r+1,\,r+1}$ is a block diagonal matrix each block of which is a tridiagonal matrix.

To solve the constructed system (1.1) we apply the generalized conjugate gradient method considered at the end of Sec. 1.2, assuming

$$A_{11}=\tilde{A}_{1,1}\oplus\ldots\oplus\tilde{A}_{r,r},\quad A_{22}=\tilde{A}_{r+1,r+1}. \tag{1.65}$$

As noted earlier, this method may be thought of as a way of optimizing (speeding up the convergence of) the block Gauss-Seidel method

$$B\mathbf{u}^k = C\mathbf{u}^{k-1} + \mathbf{f}, \tag{1.66}$$

where

$$B = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{1.67}$$

which is at the same time a particular variant of the methods using a decomposition of Ω.

It follows from the foregoing that the generalized conjugate gradient method in a subspace incorporates the following three main stages:

(i) entrance into the subspace U_A (calculation of $C\mathbf{u}^0$ and ξ^0);

(ii) implementation of $\hat{k}$ steps of the method (taking into account the structure of U_A);

(iii) going out of U_A (reconstruction of $\mathbf{u}^{\hat{k}}$ or finding a better approximation $\mathbf{v}^{\hat{k}}$).

The implementation of the first and third stages consists in sequentially solving Helmholtz difference equations in subdomains Ω_i, $i = \overline{1, r}$. Considering the assumptions made these stages can be implemented using $O(N \ln N)$ operations (in a number of special cases they can be implemented using $O(N \ln \ln N)$ operations) and when employing the methods on a sequence of grids (additional assumptions are required) using $O(N)$ arithmetic operations. Of primary interest to us is the implementation of the second stage in the subspace $U_A = \operatorname{Im} C$, where $C = B - A$.

Given for some $k \geqslant 1$ a vector $\xi^{k-1} \in U_A$, i.e. the components of the vector ξ_2^{k-1} from the representation $\xi^{k-1} = \begin{bmatrix} 0 \\ \xi_2^{k-1} \end{bmatrix}$, $\xi_2^{k-1} \in E_{N_2}$, where N_2 is the order of A_{22}, find the components of $\varphi \in E_N$ which are necessary for the last N_2 components of $CB^{-1}\xi^{k-1}$ to be calculated. It is obvious that this problem consists in finding the corresponding components of $\varphi \in E_{N_1}$ which is the solution of the system

$$A_{11}\varphi = \eta, \tag{1.68}$$

where $\eta = A_{22}A_{22}^{-1}\xi_2^{k-1}$ (calculating the components of η, which may be nonzero, requires $O(N^{1/2})$ arithmetic operations).

System (1.68) is partitioned into r subsystems

$$A_{ii}\varphi_i = \eta_i \tag{1.69}$$

which correspond to homogeneous systems of difference equations in the rectangles of Ω with homogeneous boundary conditions on

the sides parallel to the x_1-axis and with (in general) nonhomogeneous Dirichlet boundary conditions on the sides parallel to the x_2-axis. For the subsequent calculation of the components of $A_{21}\varphi$ it is only necessary to find a solution at the nodes close to the boundary and adjacent to the sides of rectangles Ω_i parallel to the x_2-axis. It is obvious that a simple solution r of such problems by the method of the separation of variables requires $O(N)$ arithmetic operations and storing at most $O(N^{1/2})$ quantities.

The stated problems for system (1.69), as in the preceding sections (they, as already noted in the introduction, relate to problems of a partial solution of systems), arise at every step of the method under consideration. It is shown in [1.34] that a multiple solution of these problems can be carried out in the following way. First in the method of the separation of variables two fast Fourier transforms are performed using $O(N^{1/2} \ln N)$ operations (since the right-hand side of system (1.69) is other than zero only on the two grid lines parallel to the x_2-axis). We then must implement a series of sweeps for the tridiagonal system, which requires $O(N)$ arithmetic operations. But the linear systems solved by the method of sweeps are very specific since on their right-hand sides only the first and last components are nonzero. Knowing the corner elements of the inverse matrices, we therefore can solve the entire set of these systems (we need only the first and last components of solutions) performing $O(N^{1/2})$ arithmetic operations. This is the essence of the suggestion that one should calculate at the first step the required $O(N^{1/2})$ corner elements of the matrices inverse to $O(N^{1/2})$ tridiagonal matrices using $O(N)$ arithmetic operations and perform at each subsequent step only $O(N^{1/2})$ operations in the procedure described. Again only two (for each rectangle Ω_i) fast Fourier transforms requiring $O(N^{1/2} \ln N)$ arithmetic operations complete the algorithm for the method of the separation of variables.

To summarize, the implementation of the first step of the method requires $O(N)$ operations, each subsequent step requiring $O(N^{1/2} \ln N)$ operations and the entire algorithm requiring only $O(N^{1/2})$ quantities to be stored.

The results of [1.34] and [1.35] yield

Lemma 1.4. *For method* (1.66) *we have the estimate*

$$\rho(B^{-1}C) \leqslant 1 - dh, \tag{1.70}$$

where $h = O(N^{-1/2})$ *and* d *is some* N*-independent constant.*

Thus we have the following

Theorem 1.5. *To solve the constructed system* (1.1) *to an accuracy of* $\varepsilon < 1$ *by the stated generalized conjugate gradient method in a subspace it suffices to perform* $O(N \ln N)$ *arithmetic operations* (*or*

$O(N \ln \ln N)$ operations) at the first and third stages and

$$O(N) + O(N^{3/4} \ln N \ln(1/\varepsilon))$$

arithmetic operations at the second stage while storing only $O(N^{1/2})$ quantities.

References

1.1. M. R. Hestenes, E. Stiefel, *J. Res. Nat. Bur. Stand.* 49: 409-436 (1952).
1.2. M. R. Hestenes, in: *Proc. Symp. Appl. Math.*, vol. 6 (Providence: Amer. Math. Soc. 1956): 82-102.
1.3. G. I. Marchuk and Yu. A. Kuznetsov, *Dokl. Akad. Nauk SSSR*, 181, No. 6: 1331-1334 (1968) (in Russian).
1.4. G. I. Marchuk and Yu. A. Kuznetsov, *Iteration Methods and Quadratic Functionals* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1972) (see also in: *Methods of Computational Mathematics* (Novosibirsk: Nauka, 1975): 4-143 (in Russian).
1.5. G. I. Marchuk and Yu. A. Kuznetsov, *The Generalized Conjugate Gradient Method: Theory and Application* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, (Preprint No. 72), 1980 (in Russian)
1.6. Yu. A. Kuznetsov, in: *Computational Processes and Computing Systems* (Moscow: Nauka, 1983): 267-300 (in Russian).
1.7. B. N. Parlett, *The Symmetric Eigenvalue Problem* (Englewood Cliffs: Prentice-Hall, 1980).
1.8. C. Lanczos, *J. Res. Nat. Bur. Stand.* 49: 33-53 (1952).
1.9. D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra* (Moscow: Fizmatgiz, 1963) (in Russian).
1.10. J. K. Reid, in: *Large Sparse Sets of Linear Equations* (London, New York: Academic Press, 1971): 231-252.
1.11. P. Concus, G. H. Golub, and D. P. O'Leary, in: *Sparse Matrix. Computations* (New York: Academic Press, 1976): 309-332.
1.12. A. A. Samarsky and E. S. Nikolaev, *Methods of Solving Grid Equations* (Moscow: Nauka, 1978) (in Russian).
1.13. V. P. Il'in and Yu. A. Kuznetsov, in: *Proc. Third Internat. Symposium on Computing Methods in Applied Sciences and Engineering, Versalle, 1977*. I, ed. J.-L. Lions and R. Glowinski (New York, Heidelberg, Berlin: Springer-Verlag (Lecture Notes in Math., No. 704), 1979): 23-36.
1.14. Yu. A. Kuznetsov, *Dokl. Akad. Nauk SSSR*, 184, No. 2: 274-277 (1969) (in Russian).
1.15. E. G. D'yakonov, in: *Numerical Methods in Mathematical Physics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 45-68 (in Russian).
1.16. G. I. Marchuk and V. I. Lebedev, *Numerical Methods in Neutron Transport Theory* (Moscow: Atomizdat, 1981) (in Russian).
1.17. Yu. A. Kuznetsov, in: *Computational Methods in Mathematical Physics, Geophysics, and Optimal Control* (Novosibirsk: Nauka, 1978): 125-137 (in Russian).
1.18. R. S. Varga, *Matrix Iterative Analysis* (Englewood Cliffs: Prentice-Hall, 1962).
1.19. Yu. A. Kuznetsov, Cand. Sc. Thesis (Phys.-Math.) (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1969) (in Russian).
1.20. Yu. A. Kuznetsov and V. I. Yavushkin, in: *Numerical Methods in Mathematical Physics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 84-105 (in Russian).
1.21. G. I. Marchuk and Yu. A. Kuznetsov, in: *Computational Methods of Linear Algebra* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1972): 4-20 (in Russian).

1.22. Yu. A. Kuznetsov and A. M. Matsokin, in: *Computational Methods of Linear Algebra* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1977): 79-86 (in Russian).
1.23. G. P. Astrakhantsev, in: Cand. Sc. Thesis (Phys.-Math.) (Leningrad: Leningrad Otdel. Matem. Inst. AN SSSR, 1972) (in Russian).
1.24. G. P. Astrakhantsev, *Zh. Vychisl. Matem. i Mat. Fiz.*, 18, No. 1: 118-125 (1978) (in Russian).
1.25. A. M. Matsokin, in: *Numerical Methods of Linear Algebra* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1973): 48-56 (in Russian).
1.26. B. L. Buzbee, J. A. Dorr, J. A. George, and G. H. Golub, *SIAM J. Numer. Anal.* 8: 722-736 (1971).
1.27. W. Proskurowski, O. Widlund, *Math. Comput.* 30: 433-468 (1976).
1.28. I. E. Kaporin and E. S. Nikolayev, *Diff. Uravn.* 16, No. 7: 1211-1225 (1980) (in Russian).
1.29. A. M. Matsokin, in: *Numerical Methods of Linear Algebra* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980): 66-77 (in Russian).
1.30. A. M. Matsokin, in: *Methods of Solving Systems of Variational Difference Equations* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 136-138 (in Russian).
1.31. A. Banegas, *Math. of Comput.* 32, No. 142: 441-446 (1978).
1.32. P. N. Swarztrauber, *SIAM Rev.* 19, No. 3: 490-501 (1977).
1.33. Yu. A. Kuznetsov and A. M. Matsokin, in: *Numerical Methods of Linear Algebra* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1978): 62-89 (in Russian).
1.34. Yu. A. Kuznetsov, in: *Variational Difference Methods in Mathematical Physics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1978): 178-212 (in Russian).
1.35. Yu. A. Kuznetsov, in: *Numerical Methods in Mathematical Physics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 20-44 (in Russian).
1.36. V. I. Lebedev and V. I. Agoshkov, *Seminar on Problems of Computational Mathematics* (supervised by Academician G. I. Marchuk) Preprint No. 19 (Moscow, VINITI, 1981) (in Russian).
1.37. V. V. Smelov, *Zh. Vychisl. Mat. i Mat. Fiz.* 21, No. 6: 1488-1504 (1981) (in Russian).

2 Problems of Projection-Grid Methods in Transport Theory

G. I. Marchuk and V. I. Agoshkov

Projection-grid methods (finite-element techniques) have been used to advantage to solve neutron transport problems, but a number of difficulties have arisen. This paper formulates some of the problems of projection-grid methods: the choice of a generalized statement of a problem, the problem of investigating the differentiability properties of the solutions of transport equations, the estimation of the rate of convergence of approximate solutions, the choice of basis functions taking into account singularities of solutions, the construction of extensions of solutions preserving the class of smoothness and their possible applications in numerical algorithms. In some cases ways of solving the problems discussed are suggested.

2.1. INTRODUCTION

Projection-grid methods (finite-element techniques) have been successfully used to solve neutron transport problems. A number of difficulties arise here, however, due to some "disagreeable singularities" of these problems, their being essentially multidimensional in the first place. Considerable difficulties associated with the justification of projection-grid methods and obtaining numerical estimates of the rate of convergence. One of the difficulties is the problem of approximating boundaries and boundary conditions, an extra problem of seeking and approximating the "lit" and "shady" parts of a boundary coming up, due to the dependence of boundary conditions on angular variables.

It should be noted that projection-grid algorithms frequently involve the function theory and the theory of boundary-value problems. Solving these questions helps eliminate many of the difficulties in projection-grid algorithms when applying these to transport problems. This paper discusses some problems of projection-grid methods. These are the choice of a generalized statement of a problem, the study of the differentiability properties of the solutions of transport equations and the estimation of the rate of convergence of approximate solutions, the choice of basis functions taking into account the singularities of solutions, and constructing the extensions of solutions preserving the class of smoothness and their pos-

sible applications in numerical algorithms. In some cases ways of solving the problems discussed are suggested.

We formulate the problems to be discussed below. Let D be a convex domain in R^3 with piecewise smooth boundary ∂D of class $C^{(1)}$, Ω be a sphere of unit radius in R^3 with centre at the origin, $\mathbf{s} = (s_1, s_2, s_3)$, $\sum_{i=1}^{3} s_i^2 = 1$, $s_1 = \sin\vartheta\cos\psi$, $s_2 = \sin\vartheta\sin\psi$, $s_3 = \cos\vartheta$ $(0 \leqslant \vartheta \leqslant \pi,\ 0 \leqslant \psi \leqslant 2\pi)$. Consider the following stationary problem for the one-velocity transport equation

$$(\mathbf{s}, \nabla)\varphi + \sigma\varphi = [\sigma_s/(4\pi)] \int_{\Omega} \Theta(\mu_0)\varphi(\mathbf{x}, \mathbf{s}')\, d\mathbf{s}' + F(\mathbf{x}, \mathbf{s}), \tag{2.1}$$

$$\varphi(\mathbf{x}, \mathbf{s}) = 0 \text{ when } \mathbf{x} \in \partial D,\ (\mathbf{s}, \mathbf{n}(\mathbf{x})) < 0, \tag{2.2}$$

where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is a unit vector of the outward normal to ∂D at a point $\mathbf{x} \in \partial D$; $(\mathbf{s}, \mathbf{n}) = \sum_{i=1}^{3} s_i n_i$; $\mu_0 = (\mathbf{s}, \mathbf{s}') = \sum_{i=1}^{3} s_i s_i'$; $(\mathbf{s}, \nabla)\varphi = \sum_{i=1}^{3} s_i \partial\varphi/\partial x_i$; $\int_{\Omega} \varphi\, d\mathbf{s} = \int_0^{2\pi} d\psi \int_0^{\pi} \sin\vartheta\varphi(\mathbf{x}, \mathbf{s})\, d\vartheta$; $\sigma(\mathbf{x})$, $\sigma_s(\mathbf{x})$ are almost everywhere bounded functions of $\mathbf{x} \in D$; $\Theta(\mu_0)$ is a nonnegative function summable over $[-1, 1]$, with $\int_{\Omega} \Theta(\mu_0)\, d\mathbf{s} = 4\pi$. It is assumed that

$$0 < a_0 \leqslant \sigma(\mathbf{x}) \leqslant a_1 < \infty,$$
$$0 \leqslant \sigma_s(\mathbf{x})/\sigma(\mathbf{x}) \leqslant b_1 < 1,\ a_0,\ a_1,\ b_1 = \text{const}.$$

Equation (2.1) will often be written as

$$(1/\sigma)(\mathbf{s}, \nabla)\varphi + \varphi = [b/(4\pi)] \int_{\Omega} \Theta(\mu_0)\varphi(\mathbf{x}, \mathbf{s}')\, d\mathbf{s}' + f(\mathbf{x}, \mathbf{s}), \tag{2.3}$$

where $b(\mathbf{x}) \equiv \sigma_s(\mathbf{x})/\sigma(\mathbf{x})$, $f(\mathbf{x}, \mathbf{s}) = F(\mathbf{x}, \mathbf{s})/\sigma(\mathbf{x})$.

We denote by B a Hilbert space in which the scalar product and the norm have the form

$$(\varphi, \psi) = \int_{\Omega \times D} \sigma(\mathbf{x})\varphi(\mathbf{x}, \mathbf{s})\psi(\mathbf{x}, \mathbf{s})\, d\mathbf{x}\, d\mathbf{s},\ \|\varphi\| = (\varphi, \varphi)^{1/2}.$$

$D(L)$ will denote a set of functions $\varphi \in B$ such that $(1/\sigma)(\mathbf{s}, \nabla)\varphi \in B$ and almost everywhere satisfying the boundary condition (2.2). Giving on $D(L)$ the operator L: $L\varphi = (1/\sigma)(\mathbf{s}, \nabla)\varphi + \varphi$ and denot-

ing by S the integral operator appearing on the right of (2.3), i.e. setting $S\varphi = [b(x)/(4\pi)] \int_\Omega \Theta(\mu_0)\,\varphi(\mathbf{x}, \mathbf{s}')\,d\mathbf{s}'$, we write the problem for equation (2.3) in operator form

$$L\varphi = S\varphi + f, \tag{2.4}$$

where it is assumed that $f \in B$. Now the problem for the transport equation is formulated as follows: find a function $\varphi \in D(L)$ satisfying equation (2.4) almost everywhere in $D \times \Omega$.

The problem conjugate to (2.4) has the form

$$L^*\psi = S\psi + g, \tag{2.5}$$

where $g \in B$ and L^* is the operator conjugate to L which is given by $L^*\psi = -(1/\sigma)(\mathbf{s}, \nabla)\psi + \psi$ and the domain of definition $D(L^*)$ consisting of functions $\psi(\mathbf{x}, \mathbf{s})$ such that $\psi(\mathbf{x}, -\mathbf{s}) \in D(L)$. Problem (2.5) can also be written in the form

$$(1/\sigma)(\mathbf{s}, \nabla)\psi + \psi = [b/(4\pi)] \int_\Omega \Theta(\mu_0)\,\psi(\mathbf{x}, \mathbf{s}')\,d\mathbf{s}' + g(\mathbf{x}, \mathbf{s}),$$
$$\psi(\mathbf{x}, \mathbf{s}) = 0 \text{ for } \mathbf{x} \in \partial D \text{ and } (\mathbf{s}, \mathbf{n}(\mathbf{x})) > 0. \tag{2.6}$$

If the domain D is a slab $D = \{\mathbf{x}: 0 \leqslant x_s \equiv z \leqslant H, -\infty \leqslant x_1, x_2 \leqslant \infty\}$, where $\sigma(\mathbf{x})$ and $\sigma_s(\mathbf{x})$ depend only on the variable z and f depends only on z and one angular variable ϑ, then the problem for equation (2.3) is written as follows:

$$(\mu/\sigma)(\partial\varphi/\partial z) + \varphi = (b(z)/2) \int_{-1}^{1} \Theta(\mu, \mu')\,\varphi(z, \mu')\,d\mu' + f(z, \mu),$$
$$\varphi(0, \mu) = 0 \text{ for } \mu > 0, \quad \varphi(H, \mu) = 0 \text{ for } \mu < 0, \tag{2.7}$$

where $\Theta(\mu, \mu') = \left[1/(2\pi)\right] \int_0^{2\pi} \Theta(\mu_0)\,d\psi$. The scalar product in B in this problem is of the form $(\varphi, \psi) = \int_0^H dz \int_{-1}^{1} \sigma(z)\,\varphi(z, \mu)\,\psi(z, \mu)\,d\mu$.

If, however, D is a sphere of radius R with centre at the origin and σ, σ_s, and f depend only on the radius $r = (x_1^2 + x_2^2 + x_3^2)^{1/2}$, with the angular dependence of f expressed by the variable $\mu = \cos\tilde{\vartheta}$, where $\tilde{\vartheta}$ is the angle between the direction $\mathbf{s}$ and the radius vector from the origin to the point $\mathbf{x}$, then in coordinates r and μ

problem (2.4) can be written as

$$\frac{1}{\sigma(r)}\left(\mu\frac{\partial\varphi}{\partial r}+\frac{1-\mu^2}{r}\frac{\partial\varphi}{\partial\mu}\right)+\varphi$$

$$=\frac{b(r)}{2}\int_{-1}^{1}\Theta(\mu,\ \mu')\,\varphi(r,\ \mu')\,d\mu'+f(r,\ \mu), \qquad (2.8)$$

$$(0<r<R,\quad -1<\mu<1),$$

$$\varphi(R,\ \mu)=0,\quad -1\leqslant\mu<0,$$

where

$$\Theta(\mu,\ \mu')=(1/\pi)\int_{0}^{\pi}\Theta(\mu,\ \mu')+[(1-\mu^2)\,(1-\mu'^2)]^{1/2}\cos\psi)\,d\psi.$$

Here the scalar product in B is

$$(\varphi,\ \psi)=\int_{0}^{R}r^2\,dr\int_{-1}^{1}\sigma(r)\,\varphi(r,\ \mu)\,\psi(r,\ \mu)\,d\mu.$$

It is easy to see that the operator of equation (2.4) is not symmetric. In a number of cases, however, it is possible to go over from the problem for equation (2.3) to the one in which all operators are symmetric, which has certain advantages. Suppose, for example, that the functions $\Theta(\mu_0)$ and $f(\mathbf{x}, \mathbf{s})$ are even with respect to angular variables, i.e. $\Theta(\mu_0)=\Theta(-\mu_0)$ and $f(\mathbf{x},\ \mathbf{s})=f(\mathbf{x},\ -\mathbf{s})$. Then it is possible to go over from the problem for equation (2.3) to the following symmetrized problem [2.1]:

$$-(1/\sigma)\,(\mathbf{s},\ \nabla)\,(1/\sigma)\,(\mathbf{s},\ \nabla)\,u+u=Su+f,$$

$$u-(1/\sigma)\,(\mathbf{s},\ \nabla)\,u=0 \text{ for } \mathbf{x}\in\partial D \text{ and } (\mathbf{s},\ \mathbf{n})<0, \qquad (2.9)$$

$$u+(1/\sigma)\,(\mathbf{s},\ \nabla)\,u=0 \text{ for } \mathbf{x}\in\partial D \text{ and } (\mathbf{s},\ \mathbf{n})>0$$

whose solution is related to the solution φ of (2.3) by

$$u(\mathbf{x},\ \mathbf{s})=(\varphi(\mathbf{x},\ \mathbf{s})+\varphi(\mathbf{x},\ -\mathbf{s}))/2,$$

$$\varphi(\mathbf{x},\ \mathbf{s})=u(\mathbf{x},\ \mathbf{s})-(1/\sigma)\,(\mathbf{s},\ \nabla)\,u(\mathbf{x},\ \mathbf{s}).$$

Problem (2.9) can be written in operator form

$$L_0u=Su+f, \qquad (2.10)$$

where

$$L_0u=-(1/\sigma)\,(\mathbf{s},\ \nabla)\,(1/\sigma)\,(\mathbf{s},\ \nabla)\,u+u,$$

$$D(L_0)=\{u:\ u\in B,\ L_0u\in B;\ u-(1/\sigma)\,(\mathbf{s},\ \nabla)\,u=0$$

$$\text{for } \mathbf{x}\in\partial D,\ (\mathbf{s},\ \mathbf{n})<0;$$

$$u+(1/\sigma)\,(\mathbf{s},\ \nabla)\,u=0 \text{ for } \mathbf{x}\in\partial D,\ (\mathbf{s},\ \mathbf{n})>0\}.$$

Note that the operator of problem (2.10) is now symmetric. Using the following formula of integration by parts [2.1]

$$(L\varphi, \psi) = \int\limits_{\Omega\times\partial D_+} (\mathbf{s}, \mathbf{n})\,\varphi(\mathbf{x}, \mathbf{s})\,\psi(\mathbf{x}, \mathbf{s})\,d\mathbf{x}\,d\mathbf{s}$$

$$- \int\limits_{\Omega\times\partial D_-} (\mathbf{s}, \mathbf{n})\,\varphi(\mathbf{x}, \mathbf{s})\,\psi(\mathbf{x}, \mathbf{s})\,d\mathbf{x}\,d\mathbf{s} + 2(\varphi, \psi) - (\varphi, L\psi), \qquad (2.11)$$

where ∂D_+ is the part of the boundary ∂D for which $(\mathbf{s}, \mathbf{n}) > 0$ with $\mathbf{s}$ fixed and ∂D_- is the part of ∂D for which $(\mathbf{s}, \mathbf{n}) < 0$, it is not hard to prove that L and L_0 are positive definite. It may also be shown [2.1] that in the problems above the operator L^{-1} acting from B into B exists and is bounded and that equation (2.4) has a unique solution φ, with $\| (\mathbf{s}, \nabla)\varphi \| + \| \varphi \| \leqslant c \| f \|$, $c = \text{const} > 0$.

2.2. THE CHOICE OF A GENERALIZED STATEMENT OF THE PROBLEM

Generalized statements of problems not only bear closely on the study of solvability of equations but they also underlie projection-grid algorithms. The choice of statement determines the requirements for the smoothness of basis functions, the necessity for satisfying boundary conditions, etc. Therefore we consider here some of the generalized statements of problems for the transport equation that are frequently used in constructing approximate solutions.

Consider Hilbert spaces

$$H_A = \{u: (u, v)_A = (u, v) + \int\limits_{\partial D\times\Omega} |(\mathbf{s}, \mathbf{n})|\,uv\,d\mathbf{x}\,d\mathbf{s}, \quad \| u \|_A = (u, u)_A^{1/2}\},$$

$$H_1 = \Big\{u: (u, v)_1 = ((1/\sigma)(\mathbf{s}, \nabla)u, (1/\sigma)(\mathbf{s}, \nabla)v)) + (u, v) + \int\limits_{\partial D\times\Omega} |(\mathbf{s}, \mathbf{n})|\,uv\,d\mathbf{x}\,d\mathbf{s}, \quad \| u \|_1 = (u, u)_1^{1/2}\Big\}.$$

Let φ be a solution of problem (2.3), (2.2) and let ψ be an arbitrary function in H_1. Perform a scalar multiplication of (2.3) in B by ψ and integrate by parts taking into account the boundary conditions. Then we obtain the equation

$$-(\varphi(1/\sigma)(\mathbf{s}, \nabla)\psi) + (\varphi, \psi) + \int\limits_{\partial D_+\times\Omega} |(\mathbf{s}, \mathbf{n})|\,\varphi\psi\,d\mathbf{x}\,d\mathbf{s} - (S\varphi, \psi) = (f, \psi) \qquad (2.12)$$

on which the definition of a generalized solution is based.

Definition. 2.1. *A generalized solution of problem* (2.3), (2.2) *is the function* $\varphi \in H_A$ *satisfying* (2.12) *for an arbitrary element* $\psi \in H_1$.

In what follows no questions of solvability of problems in generalized formulations are considered, since problem (2.3), (2.2) itself has a unique solution. This gives an answer to the question about the existence of a generalized solution. We shall be concerned with the requirements imposed on smoothness and satisfying boundary conditions in a generalized formulation as factors influencing to a considerable degree the course of a projection-grid algorithm. For example, it is not required in (2.12) that φ should satisfy the boundary condition (2.2). However, if the generalized solution $\varphi \in H_A$ possesses some additional smoothness (with nothing assumed about satisfying boundary conditions), then it turns out that $\varphi = 0$ for $\mathbf{x} \in \partial D$, $(\mathbf{s}, \mathbf{n}) < 0$. Indeed, let $\varphi \in H_1$. Then integrating by parts in (2.12) we arrive at

$$((1/\sigma)(\mathbf{s}, \nabla)\varphi + \varphi - S\varphi - f, \psi) + \int\limits_{\partial D_- \times \Omega} |(\mathbf{s}, \mathbf{n})| \varphi\psi \, d\mathbf{x} \, d\mathbf{s} = 0.$$

It follows from the arbitrariness of ψ that φ almost everywhere satisfies both equation (2.3) and the boundary condition (2.2). Thus (2.2) proves to be natural in the problem under consideration. This formulation of the problem is therefore convenient where the configuration of the domain D is such that it turns out to be sufficiently difficult to construct basis functions to satisfy boundary conditions.

Attractive is a modification of equation (2.12) which would contain no term $\int\limits_{\partial D_+ \times \Omega} |(\mathbf{s}, \mathbf{n})| \varphi\psi \, d\mathbf{x} \, d\mathbf{s}$. When implementing a projection-grid algorithm this would save us from computing the integrals over the set $\partial D_+ \times \Omega = \{(\mathbf{x}, \mathbf{s}): \mathbf{x} \in \partial D, \mathbf{s} \in \Omega, (\mathbf{s}, \mathbf{n}) > 0\}$. It is possible to have this term omitted by requiring that $\psi = 0$ for $\mathbf{x} \in \partial D$, $(\mathbf{s}, \mathbf{n}) > 0$. We thus arrive at the following statement of the problem.

Definition 2.2. *A generalized solution of problem* (2.3), (2.2) *is the function* $\varphi \in B$ *satisfying*

$$-(\varphi (1/\sigma)(\mathbf{s}, \nabla)\psi) + (\varphi, \psi) = (S\varphi, \psi) + (f, \psi) \quad (2.13)$$

for an arbitrary function $\psi \in H_1$ *such that* $\psi = 0$ *for* $\mathbf{x} \in \partial D$, $(\mathbf{s}, \mathbf{n}) > 0$.

Again it is not required that φ should satisfy boundary conditions. If it turns out that $\varphi \in H_1$, then integrating by parts in (2.13) we arrive at

$$(L\varphi - S\varphi - f, \psi) + \int\limits_{\partial D_- \times \Omega} |(\mathbf{s}, \mathbf{n})| \varphi\psi \, d\mathbf{x} \, d\mathbf{s} = 0.$$

By virtue of the choice of ψ being arbitrary the function φ again satisfies almost everywhere problem (2.3), (2.2).

When considering (2.13), however, we again come to the problem

ol satisfying the boundary conditions for ψ, but similar constraints on φ are as before removed. It is obvious that in the projection-grid method this will affect the requirements to the basis used to approximate φ and to the functions on which equation (2.3) is to be projected.

It is natural to assume that if the basis functions $\{\omega_i\}$ used to approximate the solution of the problem, φ, are in the domain of the operator of the problem, then the rate of convergence of approximate solutions to the exact one will be greater. Therefore where one succeeds in constructing $\{\omega_i\}$ so that $\{\omega_i\} \subset D(L)$ the following statement of the problem (in fact equivalent to (2.3), (2.2)) is preferable.

Definition 2.3. *A generalized solution of problem* (2.3), (2.2) *is the function* $\varphi \in H_1$, $\varphi = 0$ *for* $\mathbf{x} \in \partial D$, $(\mathbf{s}, \mathbf{n}) < 0$, *satisfying*

$$((1/\sigma)(\mathbf{s}, \nabla)\varphi, \psi) + (\varphi, \psi) = (S\varphi, \psi) + (f, \psi) \quad (2.14)$$

for an arbitrary function $\psi \in B$.

This statement of the problem is convenient in applying the Bubnov-Galerkin method, the method of least squares, and the method of moments. There is a sufficient arbitrariness in the choice of functions ψ here. If it is required that φ and ψ should be in the same set, then it is not hard to obtain different a priori estimates for φ (and for the approximate solution in the projection-grid method) and to prove that the matrix of the system is positive definite.

In (2.14) it may also be assumed that $\psi \in H_1$ or $\psi \in D(L^*)$. This will result in a modification of the projection-grid algorithm. The presence of two different bases, however, i.e. the case where φ and ψ in (2.14) are in different sets, often complicates investigating the algorithm for solving the problem, although its numerical implementation may greatly simplify (calculating matrix elements, the right-hand sides, boundary integrals, etc.).

Our next formulation of the problem refers to the case where $\Phi(\mu_0)$ and $f(\mathbf{x}, \mathbf{s})$ in (2.3) are even. Then problem (2.3), (2.2) reduces to problem (2.9) for which now a variational formulation is possible. We give it.

We introduce on $D(L_0)$ the scalar product $[u, v] = (L_0 u, v)$ which can be written as

$$[u, v] = \int_{\Omega \times \partial D} |(\mathbf{s}, \mathbf{n})| u(\mathbf{x}, \mathbf{s}) v(\mathbf{x}, \mathbf{s})\, d\mathbf{x}\, d\mathbf{s}$$
$$+ (u, v) + ((1/\sigma)(\mathbf{s}, \nabla) u, (1/\sigma)(\mathbf{s}, \nabla) v) - (Su, v) \quad (2.15)$$

and the norm $[u] = [u, u]^{1/2}$. Completing $D(L_0)$ according to the norm $[\cdot]$ we obtain the energy space B_A. (Taking into account the initial data of problem (2.3), (2.2) we notice that the norms $\|u\|_1$ and $[u]$ for $u \in B_A$ are equivalent.) Now problem (2.9) reduces to

finding a minimum of the quadratic functional

$$F(v) = [v]^2 - 2(f, v), \tag{2.16}$$

where $v \in B_A$.

Definition 2.4. *A generalized solution of problem* (2.3), (2.2) *is the function* $u \in B_A$ *realizing a minimum of* $F(v)$:

$$F(u) = \inf_{v \in B_A} F(v). \tag{2.17}$$

By virtue of the properties of the operator $L_0 - S$ the problem of minimizing (2.16) has a unique solution $u \in B_A$. The boundary conditions in (2.9) are natural here. Therefore if the problem of minimizing $F(u)$ is solved by the projection-grid method, then the basis functions may be chosen without having to satisfy the boundary conditions in (2.9). If u is a solution of the variational problem (2.17), then a generalized solution φ of problem (2.3), (2.2) is taken to be the function defined by the relation $\varphi = u - (1/\sigma)(\mathbf{s}, \nabla) u$.

Thus when choosing a generalized statement of the problem for the transport equation we must take into account a number of factors since it may strongly influence the entire process of solving the problem.

2.3. DIFFERENTIABILITY PROPERTIES OF SOLUTIONS AND ESTIMATION OF THE RATE OF CONVERGENCE

Of great importance in the theory of projection-grid methods (as well as in approximation algorithms generally) is smoothness of the solution of the problem. Obtaining this information in problems of transport theory involves great difficulties, however. Consequently, the study of smoothness properties of solutions and of estimation of the rate of convergence on a concrete class of solutions is one of the problems in the theory of approximate methods for transport problems.

However, if a set problem has a unique solution, then it is possible as a rule to select a particular smooth solution which would correspond to special initial data (it is here possible to use the well-known technique of substituting a smooth function into the equation and finding the corresponding right-hand sides, etc.). In such a case estimation of the rate of convergence of error may be carried out for that particular smooth solution. It is natural that the results obtained here cannot be carried over to a general case. But if they are sufficiently effective, one would expect the algorithm to be efficient in concrete situations as well. The most desirable result, of course, is obtained under practicable assumptions about the smoothness of the exact solu-

tion. The situation is such, however, that solutions of transport theory problems usually have no smoothness in space and angular variables, even if the coefficients and functions of the sources are sufficiently smooth. For example, the solution of the simplest problem

$$\mu\partial\varphi/\partial z + \varphi = 1,\ \varphi(0, \mu) = 0,\ \mu > 0,\ \varphi(H, \mu) = 0,$$
$$\mu < 0,$$

which is a special case of problem (2.7), can be written in explicit form and it turns out that the solution $\varphi(z, \mu)$ has no derivative $\partial\varphi/\partial z$ in $L_2([0, H] \times [-1, 1])$ (although $\varphi \in L_2$, $\mu\partial\varphi/\partial z \in L_2$).

Thus, in considering problems for the transport equation even given smooth coefficients and function f the assumption about the existence of derivatives with respect to space and angular variables of solutions in B in general is not true.

What smoothness (apart from the smoothness along the direction of the flight of particles) does a solution of a given problem possess? It turns out that $\varphi(z, \mu)$ may have smoothness "in the mean". For instance, if we consider problem (2.7) with $\Theta = 1$, $f = f(z) \in L_\infty(0, H)$, then for any $z, z' \in (0, H)$, $|z - z'| \leqslant h$ we have [2.2]:

$$\left(\int_{-1}^{1} |\varphi(z, \mu) - \varphi(z', \mu)|^p d\mu\right)^{1/p}$$
$$\leqslant c \|f\|_{L_\infty} \cdot \begin{cases} h^{1/p}, & 1 < p < \infty, \\ h(1 + |\ln h|), & p = 1, \end{cases} \tag{2.18}$$

$$|\varphi_0(z) - \varphi_0(z')| \leqslant c \|f\|_{L_\infty} \cdot h(1 + |\ln h|), \tag{2.19}$$

where

$$\varphi_0(z) = (1/2) \int_{-1}^{1} \varphi(z, \mu) d\mu.$$

Denote by $C_I^{(\alpha)}$ the set of functions of z such that in each of the intervals $(\tilde{z}_{i-1}, \tilde{z}_i)$, $i = 1, \ldots, I + 1$, $\tilde{z}_0 = 0$, $\tilde{z}_{I+1} = H$, $\tilde{z}_{i-1} < \tilde{z}_i$, they belong to the class $C^{(\alpha)}((\tilde{z}_{i-1}, \tilde{z}_i))$, $0 < \alpha < 1$, i.e. the property $|\psi(z) - \psi(z')| \leqslant c|z - z'|^\alpha$, $z, z' \in (\tilde{z}_{i-1}, \tilde{z}_i)$ holds for them. Functions $\psi \in C_I^{(\alpha)}$ may have discontinuities of the first kind at points z_i. Suppose that in (2.7) $\Theta = 1$, $f = f(z)$, the coefficients of $a(z)$ and $b(z)$ belong to $C_I^{(\alpha)}$ and that

$$|f(z) - f(z')| \leqslant c|z - z'|^\alpha,\ z, z' \in (\tilde{z}_{i-1}, \tilde{z}_i).$$

Then it can be shown that the function of integral flux $\varphi_0(z)$ has a derivative $d\varphi_0/dz$ and that

$$\left(\int_0^H |\, d\varphi_0/dz\,|^p\, dz\right)^{1/p} \leqslant c, \quad 1 \leqslant p < \infty. \tag{2.20}$$

If, however, σ, b, $f \in C_I^{(1)}$, then (assuming $b \equiv 0$ and $f \equiv 0$ outside $[0, H]$)

$$\varphi(z, \mu) = \begin{cases} G(z) + \sum\limits_{j=0}^{i} [G]_j \exp\left\{-\int\limits_{z_j}^{z} \frac{\sigma(\xi)}{\mu}\, d\xi\right\} \\ \quad - \int\limits_0^z \widetilde{\frac{\partial G}{\partial z'}}(z') \exp\left\{-\int\limits_{z'}^{z} \frac{\sigma(\xi)}{\mu}\, d\xi\right\} dz' \text{ for} \\ \qquad\qquad z \in (\tilde{z}_i, \tilde{z}_{i+1}), \ \mu > 0 \\ G(z) - \sum\limits_{j=i+1}^{I+1} [G]_j \exp\left\{\int\limits_{\tilde{z}_j}^{z} \frac{\sigma(\xi)}{\mu}\, d\xi\right\} \\ \quad + \int\limits_z^H \widetilde{\frac{\partial G}{\partial z'}}(z') \exp\left\{\int\limits_z^{z'} \frac{\sigma(\xi)}{\mu}\, d\xi\right\} dz' \text{ for} \\ \qquad\qquad z \in (\tilde{z}_i, \tilde{z}_{i+1}), \ \mu < 0, \end{cases} \tag{2.21}$$

where $G(z) = b(z)\,\varphi_0(z) + f(z)$, $[G]_k = G(\tilde{z}_k - 0) - G(\tilde{z}_k + 0) = [b]_k \varphi_0(\tilde{z}_k) + [f]_k$ and the function

$$\widetilde{\frac{\partial G}{\partial z}}(z) = \left\{\frac{\partial G}{\partial z}(z) \quad \text{for} \quad z \in (\tilde{z}_{i-1}, \tilde{z}_i), \quad i = 1, \ldots, I+1\right\}$$

may have discontinuities at the points $\tilde{z}_i$, $i = 0, 1, \ldots, I+1$.

Under the assumptions made the derivative of $\varphi_0(z)$ is of the form

$$d\varphi_0(z)/dz = (\sigma(z)/2)\,(V_1(z) + V_2(z)).$$

Here the function $V_1(z)$ has within the interval $(\tilde{z}_{i-1}, \tilde{z}_i)$ a derivative dV_1/dz for which in the neighbourhood of $\tilde{z}_i$ there is an asymptotic expression

$$dV_1(z)/dz = [\sigma(z)\,[G]_i/(z - \tilde{z}_i)\,(1 + O(|\, z - \tilde{z}_i\,|)).$$

The function $V_2(z)$ satisfies, for any $z, z' \in (\tilde{z}_{i-1}, \tilde{z}_i)$, the relation

$$|\, V_2(z) - V_2(z')\,| \leqslant c\,|\, z - z'\,|\,(1 + |\ln|\, z - z'\,|\,|^2).$$

For $\mu = 0$ the solution $\varphi(z, \mu)$ may have discontinuities at $\tilde{z}_i$, $i = 1, \ldots, I$, with $\varphi(\tilde{z}_i - 0) - \varphi(\tilde{z}_i + 0) = [b]_i \varphi_0(\tilde{z}_i) + [f]_i$. The solution $\varphi(z, \mu)$ will be continuous in the domain $(0, H) \times (-1, 1)$ if it is assumed that $b(z)$ and $f(z)$ are continuous functions. For $\varphi(z, \mu)$ to have a derivative $\partial\varphi(z, \mu)/\partial z$ in B it is not only necessary for $\sigma(z)$, $b(z)$, and $f(z)$ to be continuous but it must also be assumed that they belong to the class $C_I^{(1)}$ and that $b(+0) = b(H - 0) = f(+0) = f(H - 0) = 0$. (These last requirements tend to fail in physical problems.)

In [2.3] conditions are investigated under which the solution of problem (2.7) is in the space $H_{\alpha,1}(X)$, $X = (0, H) \times (-1, 1)$ where the norm is of the form

$$\|u\|_{H_{\alpha,1}(X)} = \left(\|u\|^2_{H_{\alpha,1}(X)} + \int_0^\infty (\|\Delta_h u\|^2_{H_{0,1}(X_h)}/h^{1+2\alpha})\, dh\right)^{1/2}, \tag{2.22}$$

where

$$\begin{aligned} &0 < \alpha < 1, \quad X_h = (0, H - h) \times (-1, 1), \\ &\Delta_h u = u(x + h, \mu) - u(x, \mu) \\ &\|u\|^2_{H_{0,1}(X)} = \sum_{m=0}^{1} \|(\mu\partial/\partial z)^m u\|^2_{L_s(X)}. \end{aligned} \tag{2.23}$$

(Here for $h \geqslant H$ the domain X_h is considered to be an empty set and in this case the integral over the set is identically zero.) The subset of functions in $H_{\alpha,1}$ that are independent of the angular variable μ coincides with the space $W_2^{1+\alpha}(0, H)$ with the fractional index $1 + \alpha$.

It has been proved that if $\sigma(z)$, $\sigma_s(z) \in L_\infty(0, H) \cap W_2^\alpha(0, H)$, $f(z, \mu) \in L_\infty(X) \cap H_{\alpha,0}(X)$, then the solution φ of problem (2.7) is in the set $H_{\alpha,1}(X) \cap L_\infty(X)$, $\mu\partial\varphi/\partial z \in H_{\alpha,0}(X) \cap L_\infty(X)$, with

$$\begin{aligned} \|\varphi\|_{A(X)} \equiv \|\varphi\|_{H_{\alpha,1}(X)} + \sum_{m=0}^{1} \|(\mu\partial/\partial z)^m \varphi\|_{L_\infty(X)} \\ \leqslant c\,[\|f\|_{H_{\alpha,0}(X)} + \|f\|_{L_\infty(X)}(1 + \|\sigma\|_{W_2^\alpha(0,H)} \\ + \|\sigma_s\|_{W_2^\alpha(0,H)})], \end{aligned} \tag{2.24}$$

where the constant c is independent of φ.

The smoothness properties of solutions of problem (2.7) and of problems for the transport equation in other geometries are studied in [2.4]-[2.12]. It is pointed out there, in particular, that a solution

of the transport equation may turn out to be discontinuous even in those subdomains of D in which the coefficients of the problem are themselves smooth. The behaviour of solutions in these subdomains is strongly influenced by the smoothness of the coefficients in the other parts of D, the smoothness and form of the boundary ∂D, of the lines of discontinuity of the initial data, etc. These and the above facts show the specific character of transport equations and it may be expected that the study of algorithms for approximate solution of these equations will turn out to be a more complicated matter than, for example, the investigation of elliptical problems and that in order to obtain estimates of the rate of convergence in problems of transport theory it will be necessary to carry out additional studies (concerning the properties of solutions, the procedure of establishing the estimates, etc.).

Let us estimate the rate of convergence of some projection-grid algorithms. First consider problem (2.9) in slab geometry. The scalar product in H_1 and the functional $F(u)$ are here of the form

$$(u, v)_1 = \int_0^H dz \int_0^1 [(\mu^2/\sigma)(\partial u/\partial z)(\partial v/\partial z) + \sigma uv]\, d\mu + \int_0^1 \mu (uv|_{z=0} + uv|_{z=H})\, d\mu, \quad (2.25)$$

$$F(u) = \int_0^H dz \int_0^1 [(\mu^2/\sigma)(\partial u/\partial z)^2 + \sigma(z) u^2(z, \mu)]\, d\mu + \int_0^1 \mu (u^2(0, \mu) + u^2(H, \mu))\, d\mu - \int_0^H dz \int_0^1 d\mu \int_0^1 b(z) u(z, \mu) u(z, \mu') \Theta(\mu, \mu')\, d\mu' - 2 \int_0^H dz \int_0^1 \sigma(z) u(z, \mu) f(z, \mu)\, d\mu. \quad (2.26)$$

Suppose that the coefficients of $\sigma(z)$ and $b(z)$ are piecewise constant with discontinuities at a finite number of points.

The approximate solution is constructed using piecewise linear functions in the variable z. To do this we introduce a grid $0 = z_0 < \ldots < z_N = H$, $h_i = z_i - z_{i-1}$ so that the discontinuities of the functions σ and b should coincide with some of the grid nodes. Each node z_i is assigned a function $\varphi_i(z)$ equal to unity at z_i, to zero at all the other nodes, and linear in each interval $(z_{j-1},$

μ). The approximate solution of the problem is sought in the form

$$u_h(z, \mu) = \sum_{i=0}^{N} a_i(\mu)\, \varphi_i(z), \tag{2.27}$$

where functions $a_i(\mu)$ are found by minimizing the functional $F(u_h)$. As a result we arrive at the system of integral equations

$$\hat{A}\mathbf{a} = \int_0^1 \Theta(\mu, \mu')\, \hat{B}\mathbf{a}(\mu')\, d\mu' + \mathbf{f}, \tag{2.28}$$

where $\hat{A}(\mu) = (A_{ij}(\mu))$, $\hat{B} = (B_{ij})$, with

$$A_{ij}(\mu) = \mu^2 \int_0^H (1/\sigma(z))\, (\partial\varphi_i/\partial z)\, (\partial\varphi_j/\partial z)\, dz + \int_0^H \sigma(z)\, \varphi_i(z)\, \varphi_j(z)\, dz$$
$$+ \mu(\varphi_i(0)\, \varphi_j(0) + \varphi_i(H)\, \varphi_j(H)),$$

$$B_{ij} = \int_0^H b(z)\, \sigma(z)\, \varphi_i(z)\, \varphi_j(z)\, dz,$$

$$\mathbf{f}(\mu) = (f_0(\mu), \ldots, f_N(\mu))^T, \quad \mathbf{f}_i(\mu) = \int_0^H \sigma(z)\, f(z, \mu)\, \varphi_i(z)\, dz,$$

$$\mathbf{a} = (a_0(\mu), \ldots, a_N(\mu))^T, \quad i, j = 0, 1, \ldots, N.$$

The fact that u_h converges to u is easy to establish by using the density of a set of functions of the form (2.27) (as $h \to 0$) in H_1. We estimate the rate of convergence of u_h to u. If the exact solution of the problem at hand had second derivatives with respect to z summable quadratically, then the approximating properties of $\{\varphi_i(z)\}$ would imply that $[u - u_h] \leqslant O(h)$. As noted above, however, the assumption that $u(z, \mu)$ has such a smoothness fails even in the case of constant coefficients. In the problem on hand, using propositions (2.18) and the equations $((1/\sigma)\, \mu\partial u/\partial z, (1/\sigma)\, \mu\partial\varphi_i/\partial z) = ((\mu/\sigma)\, \partial u_I/\partial z, (\mu/\sigma)\, \partial\varphi_i/\partial z)$, $u_I = \sum_{i=0}^{N} u(z_i, \mu)\, \varphi_i(z)$ and considering the formulated algorithm as projection form of the method of integral identities [2.13], [2.14] we succeed in proving that for $f \in L_\infty((0, H) \times (-1, 1))$ we have an error estimate

$$(\|(\mu/\sigma)\, \partial(u - u_I)/\partial z\|^2 + \|u - u_h\|^2 + \int_0^1 \mu((u_I - u_h)^2|_{z=0}$$
$$+ (u_I - u_h)^2|_{z=H})\, d\mu)^{1/2} \leqslant c h^{1/2} \|f\|_{L_\infty}, \quad h = \max_i h_i, \tag{2.29}$$

whose left-hand side differs from the norm of error arising in the Ritz method.

We now consider the Bubnov-Galerkin method as applied to the solution of problem (2.8). As a generalized statement we take the statement in Definition 1 (see equation (2.12)). We introduce in $[0, R]$ and $[-1, 1]$ the grids $r_j = jh_r$, $h_r = R/N$, $\mu_i = ih_\mu$, $h_\mu = 2/M$, respectively, construct piecewise linear "hill functions" $\varphi_i(r)$, $i = 0, \ldots, N$, $\varphi_j(\mu)$, $j = 0, \ldots, M$, $\varphi_i(r_j) = \delta_{ij}$, $\varphi_j(\mu_r) = \delta_{jr}$, and form bilinear functions $\psi_{ij}(r, \mu) = \varphi_i(r)\varphi_j(\mu)/\gamma_{ij}$, $i = 0, \ldots, N$, $j = 0, \ldots, M$, where $\gamma_{ij} = (\beta_i h_r h_\mu)^{-1/2}$, $\beta_0 = 0.1r_1^2$, $\beta_1 = 0.5r_1^2$, $\beta_i = r_i^2$, $i = 2, \ldots, N$. The approximate solution φ_h will be sought as

$$\varphi_h(r, \mu) = \sum_{i=0}^{N} \sum_{j=0}^{M} a_{ij}\psi_{ij}(r, \mu)$$

and a_{ij} will be determined from the system of equations

$$A(\varphi_h, \psi_{ij}) = (f, \psi_{ij}), \quad i = 0, \ldots, N, \quad j = 0, \ldots, M, \tag{2.30}$$

or in matrix form

$$\hat{A}\mathbf{a} = \mathbf{f}, \tag{2.31}$$

where $\mathbf{f} = (f_{ij})$, $\mathbf{a} = (a_{ij})$, $\hat{A} = (A_{ijlk})$, $A_{ijlk} = A(\psi_{ij}, \psi_{lk})$, $l, i = 0, \ldots, N$, $k, j = 0, \ldots, M$, with

$$A(\varphi, \psi) = \int_0^1 \mu R^2 \varphi(R, \mu)\psi(R, \mu)\, d\mu$$

$$+ \int_0^R r^2\, dr \int_{-1}^{1} \left[\sigma\varphi\psi - \varphi\left(\mu\, \partial\psi/\partial r + \frac{1-\mu^2}{r}\frac{\partial\psi}{\partial\mu}\right)\right] d\mu$$

$$- \int_0^R r^2\, dr \int_{-1}^{1} [\sigma(r)\, b(r)/2] \int_{-1}^{1} \Theta(\mu, \mu')\varphi(r, \mu')\, d\mu' \psi(r, \mu)\, d\mu.$$

It is easy to prove in this algorithm that the matrix $\hat{A}$ is positive definite. Hence system (2.31) has a unique solution. In addition, since the function, $\{\psi_{ij}\}$ are uniformly linearly independent, i.e.

$$c\|\mathbf{b}\|_2 \leqslant \Big\| \sum_{i=0}^{N} \sum_{j=0}^{M} b_{ij}\psi_{ij} \Big\|_{L_2(D\times[-1, 1])}$$

$$\leqslant \tilde{c}\|\mathbf{b}\|_2 \equiv \tilde{c}\Big(\sum_{i=0}^{N} \sum_{j=0}^{M} b_{ij}^2\Big)^{1/2},$$

where c and $\tilde{c}$ are positive constants independent of $\{b_{ij}\}$, N and M, the algorithm will be also numerically stable [2.15].

We investigate the problem of approximation and convergence in the projection-grid algorithm described. Using the usual arguments [2.13] we arrive at the relations

$$\| \varphi - \varphi_h \|_A \leqslant c \inf_{\{c_{ij}\}} \| \varphi - \sum_{i=0}^{N} \sum_{j=0}^{M} c_{ij} \psi_{ij}(r, \mu) \|_1$$

$$\leqslant \varepsilon(h) \to 0, \quad h = \max(h_r, h_\mu) \to 0. \qquad (2.32)$$

To obtain estimates of the rate of convergence $\varepsilon(h)$ in (2.32) in explicit form it is necessary to use the known results on the smoothness of the solution φ. And here, too, as already noted, solutions of transport problems fail as a rule to have classical smoothness in real situations. This smoothness may be assumed, however, for some specially selected particular solution and the function of sources. Using particular basis functions $\{\psi_{ij}\}$ it is then possible to find the estimate $\varepsilon(h)$ in explicit form.

Suppose, for example, that in (2.32) the function $\varphi(r, \mu)$ is twice continuously differentiable and $\{\psi_{ij}(r, \mu)\}$ is a basis constructed using the tensor products of one-dimensional hill functions. Then, as is known, $\varepsilon(h) \leqslant ch \| \varphi \|_{C^{(2)}}$ and $\| \varphi - \varphi_h \|_A \leqslant ch \| \varphi \|_{C^{(2)}}$. Assume that we are interested in the error estimate not in the metric $\| \cdot \|_A$ but in a weaker metric, for example, in $\| \cdot \|$. The following technique is often used in elliptical problems to obtain such an estimate. An auxiliary problem $L^*\Phi = S\Phi + G$, where $G = \varphi - \varphi_h$ is considered. Subsequently Φ is again approximated by basis functions. As a result we obtain an increased rate of convergence of φ_h to φ as $h \to 0$ in the metric $\| \cdot \|$ provided the estimate for the approximation $\| \Phi - \sum_i \sum_j c_{ij} \psi_{ij} \|_1 < ch \| \Phi \|_{C^{(2)}} \leqslant ch \| G \|$ is valid. But this last estimate tends to fail (even if φ was assumed to be smooth). The trouble here is that the right-hand side of the equation $L^*\Phi = S\Phi + G$ is a priori set equal to $\varphi - \varphi_h$ and we must no longer assume that Φ is smooth but instead restrict ourselves to the available smoothness. And at this point we are no longer sure that $\Phi \in C^{(2)}$. In addition, even if we succeed in proving that $\Phi \in C^{(2)}$ the estimate $\| \Phi \|_{C^{(2)}} \leqslant c \| G \|$ often fails to hold. Therefore the method successfully employed in elliptical problems for investigating the rate of convergence in weaker metrics is often inefficient in solving transport problems by the Ritz or Bubnov-Galerkin methods.

One of the ways to be suggested for obtaining an estimate of the rate of convergence in $\| \cdot \|$ when considering problem (2.1), (2.2) with a smooth exact solution is that of directly obtaining that esti-

mate without considering any auxiliary problems. It is natural that it will require a more detailed study of approximation errors than in the usual projection-grid algorithms. This method was used in [2.15] in studying the rate of convergence of approximate solutions $\varphi_h(r, \mu)$ in considering problem (2.8) whose solution was assumed sufficiently smooth. The reasoning proceeded as follows. The interpolant was introduced

$$\varphi_I(r, \mu) = \sum_{i=0}^{N} \sum_{j=0}^{M} \gamma_{ij} \varphi(r_i, \mu_j) \psi_{ij}(r, \mu) \tag{2.33}$$

and the equations

$$A(\varphi_I, \psi_{ij}) = (f, \psi_{ij}) + \varepsilon_{ij} \tag{2.34}$$

were written with $\varepsilon_{ij} = A(\varphi_I - \varphi, \psi_{ij})$, $i = 0, \ldots, N$, $j = 0, \ldots, M$. The following assertion about the approximation was then proved: for any set of constants $\{b_{ij}\}$ and $v_h(r, \mu) = \sum_{i=0}^{N} \sum_{j=0}^{M} b_{ij} \psi_{ij}(r, \mu)$ the inequality

$$\left| \sum_{i=0}^{N} \sum_{j=0}^{M} \varepsilon_{ij} b_{ij} \right| \leqslant O(h_r^2 + h_\mu^2) \| v_h \|_A^2 \tag{2.35}$$

is true. A system of algebraic equations

$$A(\xi_h, \psi_{ij}) = \varepsilon_{ij}, \quad i = 0, \ldots, N, \quad j = 0, \ldots, M,$$

was then obtained from (2.30) and (2.34) for $\xi_h = \varphi_I - \varphi_h = \sum_{i,j} c_{ij} \psi_{ij}$, where $c_{ij} = \gamma_{ij} \varphi(r_i, \mu_j) - a_{ij}$. Estimating the vector of the right-hand side of that system and using (2.35) and $\| \varphi - \varphi_I \|_A \leqslant O(h_r^2 + h_\mu^2)$ yielded the estimate of the rate of convergence

$$\left(R^2 \int_{-1}^{1} |\mu| \, |\varphi - \varphi_h|^2 \, d\mu + \| \varphi - \varphi_h \|^2 \right)^{1/2} \leqslant O(h_r^2 + h_\mu^2). \tag{2.36}$$

(Note that using the usual scheme of reasoning of the Bubnov-Galerkin method we would have obtained in (2.32) $\varepsilon(h) = O(h_r + h_\mu)$ and hence the value $O(h_r + h_\mu)$ on the right-hand side of estimate (2.36).)

2.4. CONSTRUCTING BASIS FUNCTIONS THAT ALLOW FOR SINGULARITIES OF A SOLUTION

Of wide repute in the theory of projection-grid algorithms is the method of introducing singular functions into a basis for the

purpose of taking into account the singularities of the solution of the problem at hand [2.16]-[2.18]. This approach is most widely used in elliptical problems in which the singularities of solutions are localized in the neighbourhood of the discontinuities of coefficients, angular points, etc. As noted earlier, solutions of transport equations have no such property of localization of singularities. Hence an attempt to allow for detected singularities of solutions of transport equations by condensing the grids or by introducing singular functions even in relatively simple problems involves great difficulties which not infrequently make the algorithm impracticable. The authors are not aware of any work suggesting versatile ways of allowing for singularities of solutions in projection-grid algorithms when applied to transport problems. It is possible, however, for some problems to use the following transport algorithms in which the basis functions are constructed so that they have basic singularities of solutions, with the form itself of singularities not required. We describe this algorithm using the example of problem (2.7) for $\Theta = 1$, $f = f(z)$ as an illustration.

Suppose that $\sigma(z)$, $b(z)$, and $f(z)$ are functions, piecewise smooth in z, of class $C^{(1)}$ with possible discontinuities of the first kind in a finite number of points $\tilde{z}_i$, $i = 1, \ldots, I$.

We introduce in $[0, H]$ a grid with nodes $0 = z_0 < z_1 < \ldots < z_{N-1} < z_N = H$ so that for $h_j = z_j - z_{j-1}$ and $h = H/N$ we have

$$0 < c_0 h \leqslant h_j \leqslant c_1 h, \quad j = 1, \ldots, N, \tag{2.37}$$

where the constants c_0 and c_1 are independent of j. We assume that some of the nodes coincide with the points $\tilde{z}_i$, $i = 1, \ldots, I$, $I < N$.

We define the simplest finite functions

$$\varphi_j(z) = \frac{1}{\sqrt{h}} \begin{cases} 1 & \text{for } z_{j-1} \leqslant z < z_j, \ j = 1, \ldots, N, \\ 0 & \text{in } (z_{k-1}, z_k), \ k \neq j \end{cases} \tag{2.38}$$

and construct functions $\{\Phi_j(z, \mu)\}$, where Φ_j is a solution of the problem $L\Phi_j = \varphi_j$, $j = 1, \ldots, N$, of the form

$$\Phi_j(z, \mu) = \begin{cases} \displaystyle\int_0^z \sigma(z')\,\varphi_j(z') \exp\left\{-\int_{z'}^z [\sigma(x)/\mu]\,dx\right\} dz'/\mu, & \mu > 0, \\ \displaystyle\int_z^H \sigma(z')\,\varphi_j(z') \exp\left\{\int_z^{z'} [\sigma(x)/\mu]\,dx\right\} dx/(-\mu), & \mu < 0. \end{cases}$$

It is obvious that $\{\Phi_j\}$ are linearly independent. They can be written in explicit form for the practically important case of piecewise constant coefficients. But if $\sigma(z)$ is piecewise smooth, then $\Phi_j(z, \mu)$ can be calculated using the simplest quadrature formulas and taking into account the finiteness of $\varphi_i(z)$.

The functions $\{\Phi_j\}$ satisfy the boundary conditions in (2.7) and have singularities of the solution of the problem that arise from the presence of the operator L. In addition, $\{\Phi_j(z, \mu)\}$ reflect the behaviour of $\varphi(z, \mu)$ as a function of μ as well (which follows from the approximation results below). Hence we need not introduce additional basis functions in the angular variable here, which considerably reduces the order of the system being solved.

An approximate solution of the problem will be sought in the form

$$\varphi_h(z, \mu) = \sum_{i=1}^{N} a_i \Phi_j(z, \mu). \tag{2.39}$$

Taking as generalized statement of problem (2.14) we determine the unknown constants $\{a_i\}$ from the system of equations

$$(L\varphi_h - S\varphi_h, \varphi_j) = (f, \varphi_j), \quad j = 1, \ldots, N, \tag{2.40}$$

where $\varphi_j = L\Phi_j$. This system has a positive definite symmetric matrix and is uniquely solvable.

We write (2.40) as

$$d_{ii}a_i = \sum_{j=1}^{N} B_{ij}a_j + f_i, \quad i = 1, \ldots, N, \tag{2.41}$$

where $d_{ii} = (\varphi_i, \varphi_i)$, $B_{ij} = (S\Phi_i, \varphi_j)$, $f_i = (f, \varphi_i)$. There is a diagonal matrix on the left of (2.41), which simplifies solution of the system by iterative methods. Further, owing to the properties of the assumed basis functions $\{\varphi_i\}$ it becomes easier to calculate the matrix elements and the values of (f, φ_i). The calculations are not difficult to carry out using the appropriate quadrature formulas. If the coefficients of the problem are piecewise constant, then d_{ii} and B_{ij} in (2.41) can be found in explicit form. For example,

$$B_{ii} = 2b_i(\sigma_i h_i + E_3(\sigma_i h_i) - 1/2),$$

where $\sigma(z) = \sigma_i$, $b(z) = b_i$, $z \in [z_{i-1}, z_i]$, $E_n(x) = \int_1^\infty (e^{-xy}/y^n)\, dy$.

The values of $E_3(x)$ and e^{-x} can be calculated accurately and economically using the methods of tabulation and expansion into series of these functions and their various approximations.

We estimate the rate of convergence of φ_h to φ as $h \to 0$. It can be shown [2.13] that

$$\| L(\varphi - \varphi_h) \| \leqslant c \inf_{\{c_i\}} \| Q(\varphi, f) - \sum_{i=1}^{N} c_i \varphi_i(z) \|, \tag{2.42}$$

where $Q(\varphi, f) = (1/2)\int_{-1}^{1} \varphi\, d\mu + f(z)$. Thus the problem of estimating the rate of convergence reduces to that of approximating the function $Q(\varphi, f)$ (which in each interval $(\tilde{z}_i, \tilde{z}_{i+1})$ is smoother than the solution $\varphi(z, \mu)$!) with the aid of the assumed finite functions $\{\varphi_i(z)\}$. Since with the step functions $\{\varphi_i\}$ we have $\inf_{\{c_i\}} \| Q(\varphi, f) - \sum_i c_i \varphi_i \| \leqslant O(h)$, we obtain an estimate (in general unimprovable)

$$\| L(\varphi - \varphi_h) \| \leqslant O(h). \tag{2.43}$$

It can also be proved that for the functions of integrated fluxes $\varphi_0(z) = \int_{-1}^{1} \varphi\, d\mu/2$, $\varphi_{0,h} = \int_{-1}^{1} \varphi_h\, d\mu/2$ we have the estimate

$$\| \varphi_0 - \varphi_{0,h} \|_{L_1(0,H)} \leqslant O(h^2 \ln(1/h)), \quad h \to 0. \tag{2.44}$$

We show the form of functions $\{\Phi_i(z, \mu)$ for the case of a piecewise constant coefficient $\sigma(z) = \{\sigma_i = \text{const for } z_{i-1} \leqslant z < z_i,$ $i = 1, \ldots, N\}$.

Case $\mu > 0$:

$$\Phi_i = \frac{1}{\sqrt{h}} \begin{cases} 0, & z \leqslant z_{i-1}; \\ 1 - \exp\left\{-\dfrac{\sigma_i (z - z_{i-1})}{\mu}\right\}, & z_{i-1} < z \leqslant z_i; \\ \exp\left\{-\dfrac{\sigma_{i+1}(z - z_i)}{\mu}\right\}\left(1 - \exp\left\{-\dfrac{\sigma_i h_i}{\mu}\right\}\right), & z_i < z \leqslant z_{i+1}; \\ \exp\left\{-\dfrac{\sigma_j (z - z_{j-1})}{\mu}\right\} - \sum\limits_{l=i+1}^{j+1} \dfrac{\sigma_l h_l}{\mu}\left(1 - \exp\left\{-\dfrac{\sigma_l h_l}{\mu}\right\}\right), & \\ \qquad z > z_{i+1},\ z_{j-1} < z < z_j, \quad j = i+2, \ldots, N. \end{cases}$$

Case $\mu < 0$:

$$\Phi_i = \frac{1}{\sqrt{h}} \begin{cases} 0, & z \geqslant z_i; \\ 1 - \exp\left\{\dfrac{\sigma_i (z_i - z)}{\mu}\right\}, & z_{i-1} \leqslant z < z_i; \\ \exp\left\{\dfrac{\sigma_{i-1}(z_{i-1} - z)}{\mu}\right\}\left(1 - \exp\left\{\dfrac{\sigma_i h_i}{\mu}\right\}\right), & \\ \qquad z_{i-2} \leqslant z < z_{i-1}; & \\ \exp\left\{\dfrac{\sigma_j (z_j - z)}{\mu}\right\} + \sum\limits_{l=j+1}^{i-1} \dfrac{\sigma_l h_l}{\mu}\left(1 - \exp\left\{\dfrac{\sigma_l h_l}{\mu}\right\}\right), & \\ \qquad z < z_{i-2}, \quad z_{j-1} \leqslant z < z_j, \quad j = i-2, \ldots, 1. \end{cases}$$

To conclude, the algorithm with basis functions of the type $\{\Phi_j\}$ can be extended to a number of other problems for the transport equation (see [2.7] and [2.13]) As shown by numerical experiments, even with a small number of basis functions $\{\Phi_j\}$ the approximate solution φ_h constructed by the algorithm described here is a sufficiently accurate approximation of the solution φ of the problem [2.13].

2.5. EXTENSION OF SOLUTIONS IN THE SAME CLASSES OF SMOOTHNESS AND THE FICTITIOUS DOMAIN METHOD IN PROJECTION-GRID ALGORITHMS

We shall discuss yet another matter closely related to the theory of approximation and the theory of projection-grid algorithms in particle transport problems. This is the problem of extending solutions of transport equations in the same classes of smoothness. By an extension of a function φ defined on some set X in a space $A(X)$ we mean a function $\tilde{\varphi}$ defined on a set $\tilde{X} \supset X$ and satisfying the conditions: (a) $\varphi = \tilde{\varphi}$ on X; (b) $\| \tilde{\varphi} \|_{A(\tilde{X})} \leqslant c \| \varphi \|_{A(X)}$, where c is independent of φ, $\tilde{\varphi}$.

Extensive results have been obtained on the construction of extensions of functions in the same classes of smoothness ([2.19], [2.20], etc.). The solution of this problem in transport problems is specific. The thing is that the order of differentiability of solutions of transport equations with respect to the direction s is by a unity greater than with respect to the directions orthogonal to s (with respect to which a solution may not have any smoothness at all). Besides, s may be any direction and since the domain D is fixed, we cannot assume that in a coordinate system in which one of the axes coincides with s the domain D is a rectangle, a parallelepiped, etc. This fact makes it impossible to use a number of statements obtained, for example, in the theory of elliptical problems. Constructing extensions of solutions in transport theory should therefore be regarded as a problem in its own right.

In [2.3] a number of statements is proved suggesting that in problems with slab geometry extensions of solutions in the same classes of smoothness can be constructed using well-known methods. Thus, suppose for simplicity that $[0, H] \subset (-\pi, \pi)$ and $\tilde{X} = (-\pi, \pi) \times (-1, 1)$. Then, if the solution $\varphi(z, \mu)$ of problem (2.7) satisfies inequality (2.24), this proves [2.3] that it can be extended to the set $\tilde{X}$ in the same class of smoothness of $A(\tilde{X})$. The extension $\tilde{\varphi}$ will be a finite function in z in the interval $(-\pi, \pi)$ and a solution of some problem of the type (2.7) on $\tilde{X}$.

In considering the problem of extending solutions in the case of

three space variables the study of many questions becomes more complicated. However, it is sometimes quite easy to construct these extensions. Consider, for example, problem (2.1), (2.2). Introduce a space $H_{\alpha,k}$ $(0<\alpha<1)$ with the norm

$$\|u\|_{H_{\alpha,k}(X)}=\left(\|u\|^2_{H_{0,k}(X)}+\sum_{m=0}^{k}\int_0^{\infty}\int_{\Omega}(\|\Delta_{hs'}u\|^2_{H_{0,m}(X_{hs'})}/h^{1+2\alpha})\,d\mathbf{s}'\,dh\right)^{1/2},$$

where $X=D\times\Omega$, $\Delta_{hs'}u=u(\mathbf{x}+h\mathbf{s}',\ \mathbf{s})-u(\mathbf{x},\ \mathbf{s})$, $X_{hs'}=D_{hs'}\times\Omega$, $\|u\|^2_{H_{0,k}(X)}=\sum_{m=0}^{k}\|(\mathbf{s},\ \nabla)^m u\|^2_{L_2(X)}$ and $D_{hs'}$ is a set of points $\mathbf{x}\in D$ such that lying together with $\mathbf{x}$ in D is any point $\mathbf{x}+h\mathbf{s}'\xi$, where $0\leqslant\xi\leqslant 1$.

Assuming that $\sigma,\ \sigma_s\in L_\infty(D)\cap W_2^{\alpha}(D)$, $f(\mathbf{x},\ \mathbf{s})\in L_\infty(X)\cap H_{\alpha,0}(X)$, the solution of problem (2.1), (2.2) is in the set $H_{\alpha,1}(X)\cap L_\infty(X)$ and we have the estimate [2.3]

$$\|\varphi\|_{A(X)}\equiv\|\varphi\|_{H_{\alpha,1}(X)}+\sum_{m=0}^{1}\|(\mathbf{s},\ \nabla)^m\varphi\|_{L_\infty(X)}$$
$$\leqslant c\left[\|f\|_{H_{\alpha,0}(X)}+\|f\|_{L_\infty(X)}\left(\|\sigma\|_{W_2^{\alpha}(D)}+\|\sigma_s\|_{W_2^{\alpha}(D)}+1\right)\right], \quad (2.45)$$

where c is a positive constant.

We consider further only the case where $\alpha\in(0,\ 1/2)$. Let $\tilde{D}$ be any parallelepiped with respect to which $\bar{D}=D+\partial D$ is a strictly interior subdomain and let $\tilde{X}=\tilde{D}\times\Omega$. We extend $\sigma(\mathbf{x})$ to $\tilde{D}\setminus D$ by any positive constant, and $\sigma_s(\mathbf{x})$ and $f(\mathbf{x},\ \mathbf{s})$ by zeros. Denote the extensions thus constructed by $\tilde{\sigma}$, $\tilde{\sigma}_s$ and $\tilde{f}$. Note that these functions are extensions of σ, σ_s, and f, respectively, in the same classes of smoothness. Now consider in $\tilde{D}\times\Omega$ a problem of the form

$$(\mathbf{s},\ \nabla)\tilde{\varphi}+\tilde{\sigma}\tilde{\varphi}=[\tilde{\sigma}_s/(4\pi)]\int_{\Omega}\Theta(\mu_0)\,\tilde{\varphi}\,d\mathbf{s}'+\tilde{f},$$
$$\tilde{\varphi}=0 \quad \text{for} \quad x\in\partial\tilde{D} \quad \text{and} \quad (\mathbf{s},\ \tilde{\mathbf{n}}(x))<0. \quad (2.46)$$

Note that $\tilde{\varphi}=0$ for $x\in\partial D$ and $(\mathbf{s},\ \mathbf{n}(x))<0$. Therefore $\tilde{\varphi}=\varphi$ in X and $\tilde{\varphi}$ is an extension of φ in the same classes of smoothness [2.3].

We illustrate possible applications of the above method of extending a solution of problem (2.1), (2.2) to the domain $(\tilde{D}\setminus D)\times\Omega$ in projection-grid algorithms. Suppose that the configuration of a (convex!) domain D with curvilinear boundary ∂D is inconvenient for some reasons for numerical solution. According to the foregoing

we can then consider problem (2.46) whose solution is now defined in both the domain $D \times \Omega$ and a "fictitious domain" $(\tilde{D} \setminus D) \times \Omega$ with $\tilde{D}$ being now a domain of simple configuration. If we now construct an approximation to $\tilde{\varphi}$ in $\tilde{D} \times \Omega$, then, considered only on $D \times \Omega$, it can be thought of as an approximation to φ as well.

To construct an approximate solution of problem (2.46) we introduce in R^3 (for simplicity) a uniform grid with spacing h in each variable x_i, $i = 1, 2, 3$. We form functions $\varphi_j(\mathbf{x}) = \prod_{i=1}^{3} \varphi_{ji}(x_i)$, where $\varphi_{ji}(x_i) = \varphi(x_i/h - j_i)/\sqrt{h}$, $j = (j_1, j_2, j_3)$, $\varphi(t)$ is a standard "hill function". An approximate solution $\tilde{\varphi}_h$ is sought in the form

$$\tilde{\varphi}_h = \sum_j a_j(\mathbf{s})\, \varphi_j(\mathbf{x}), \tag{2.47}$$

where summation is taken over those indices that correspond to functions $\varphi_j(\mathbf{x})$ with supports having nonzero intersections with $\tilde{D}$. We take a generalized statement of the problem in the form (2.12) and determine the unknowns $\{a_j(\mathbf{s})\}$ by the Bubnov-Galerkin method from the system of equations

$$-\int_{\tilde{D}} \tilde{\varphi}_h(\mathbf{s}, \nabla)\, \varphi_j\, d\mathbf{x} + \int_{\tilde{D}} (\tilde{\sigma}\tilde{\varphi}_h - [\tilde{\sigma}_s/(4\pi)] \int_{\Omega} \Theta \tilde{\varphi}_h\, d\mathbf{s}')\, \varphi_j\, d\mathbf{x}$$
$$+ \int_{\partial \tilde{D}_+} |(\mathbf{s}, \tilde{\mathbf{n}}(\mathbf{x}))| \tilde{\varphi}_h \varphi_j\, d\mathbf{x} = \int_{\tilde{D}} \tilde{\sigma} \tilde{f} \varphi_j\, d\mathbf{x}, \tag{2.48}$$

where j varies over the same range as it does in the sum (2.47). System (2.48) has a unique solution [2.13]. Its further numerical solution as a system of integral equations can be constructed by well-known methods. We therefore restrict our discussion to the estimation of error $(\tilde{\varphi} - \tilde{\varphi}_h)$. In the usual way [2.13] we arrive at the relation

$$\|\tilde{\varphi} - \tilde{\varphi}_h\|_A^2 \equiv (1/2) \int_{\partial \tilde{D} \times \Omega} |(\mathbf{s}, \tilde{\mathbf{n}}(\mathbf{x}))|\, |\tilde{\varphi} - \tilde{\varphi}_h|^2\, d\mathbf{x}\, d\mathbf{s}$$
$$+ \|\tilde{\varphi} - \tilde{\varphi}_h\|^2 \leqslant c \left\{ \|\tilde{\varphi} - \tilde{\varphi}_I\|^2_{H_{0,1}(\tilde{D} \times \Omega)} \right.$$
$$\left. + \int_{\partial \tilde{D}_+ \times \Omega} |(\mathbf{s}, \tilde{\mathbf{n}})|\, |\tilde{\varphi} - \tilde{\varphi}_I|^2\, d\mathbf{x}\, d\mathbf{s} \right\}, \tag{2.49}$$

where $\tilde{\varphi}_I = \sum_j b_j(\mathbf{s})\, \varphi_j(\mathbf{x})$ is a function with arbitrary coefficients

$\{b_j(\mathbf{s})\}$ which belong to the same class as the solution of system (2.48) does. Let $\tilde{\varphi}$ be extended in a finite manner to R^3 and denote this extension by $\tilde{\tilde{\varphi}}$. Then, by extending the limits of summation in $\tilde{\varphi}_I$ so that the sum contains only b_j of the functions φ_j whose supports have nonzero intersections with those of $\tilde{\tilde{\varphi}}$, we obtain $\tilde{\tilde{\varphi}}_I$. Hence (2.49) yields

$$\| \tilde{\varphi} - \tilde{\varphi}_h \|_A^2 \leqslant c \left\{ \| \tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I \|^2_{H_{0,1}(R^3 \times \Omega)} + \int\limits_{\partial \tilde{D} \times \Omega} |(\mathbf{s}, \tilde{\mathbf{n}})| |\tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I|^2 \, d\mathbf{x} \, d\mathbf{s}, \right. \quad (2.50)$$

where $\tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I$ is an R^3-finite function. Its support is in some convex bounded domain which will be denoted by $\tilde{\tilde{D}}$. Using now the inequality [2.3]

$$\int\limits_{\partial \tilde{D} \times \Omega} |(\mathbf{s}, \tilde{\mathbf{n}})| |\tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I|^2 \, d\mathbf{x} \, d\mathbf{s} \leqslant c \| \tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I \|^2_{H_{0,1}(\tilde{\tilde{D}} \times \Omega)},$$

we get

$$\| \tilde{\varphi} - \tilde{\varphi}_h \|_A^2 \leqslant c \| \tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I \|^2_{H_{0,1}(R^3 \times \Omega)} = c \int\limits_{\Omega} d\mathbf{s} \int\limits_{R^3} (1 + (\mathbf{s}, \mathbf{y})^2) |\hat{\tilde{\tilde{\varphi}}} - \hat{\tilde{\tilde{\varphi}}}_I)|^2 \, d\mathbf{y}, \quad (2.51)$$

where $\hat{v}$ denotes the Fourier transform of the function $v(\mathbf{x}, \mathbf{s})$ with respect to variables $\mathbf{x}$. If we now use the known results of approximation [2.16], [2.21] using functions $\{\varphi_j(\mathbf{x})\}$ and relation (2.45) it may be stated that for $\tilde{\tilde{\varphi}} \in H_{\alpha,1}(R^3 \times \Omega)$ there is a combination $\tilde{\tilde{\varphi}}_I$ such that

$$\| \tilde{\tilde{\varphi}} - \tilde{\tilde{\varphi}}_I \|^2_{H_{0,1}(R^3 \times \Omega)} \leqslant c h^{2\alpha} \| \tilde{\tilde{\varphi}} \|^2_{H_{\alpha,1}(R^3 \times \Omega)} \leqslant c h^{2\alpha} \| \varphi \|^2_{H_{\alpha,1}(D \times \Omega)}$$

$$\leqslant c h^{2\alpha} [\| f \|_{H_{\alpha,0}(D \times \Omega)} + \| f \|_{L_\infty(D \times \Omega)} (1 + \| \sigma \|_{W_2^\alpha(D)} + \| \sigma_s \|_{W_2^\alpha(D)})]^2.$$

We thus arrive at the estimate

$$\| \tilde{\varphi} - \tilde{\varphi}_h \|_A \leqslant O(h^\alpha). \quad (2.52)$$

Considering $\tilde{\varphi}_h$ only in $D \times \Omega$ as an approximate solution φ_h of problem (2.1), (2.2) we come to the estimate

$$\| \varphi - \varphi_h \|_{L_2(D\times\Omega)} \leqslant O(h^{\alpha}). \tag{2.53}$$

Remark. It is naturally possible to assume $\tilde{D} = D$ in the algorithm described (i.e. no transition to a wider domain $\tilde{D}$ is effected). Then the stronger estimate (2.52) is also valid for the approximate solution.

Note also that the above algorithm (which may also be regarded as a combination of the fictitious domain method and the projection-grid algorithm) suggests that for numerical solution of many transport problems in convex domains with curvilinear boundaries it suffices to develop efficient methods for solving such problems in parallelepipeds. And in considering these domains it is very convenient to use the various splitting methods [2.13], [2.14], [2.22] which frequently allow one to reduce the original multidimensional problem to a serial solution of one-dimensional problems.

References

2.1. V. S. Vladimirov, *Trudy MIAN SSSR* 61: 3-158 (1961) (in Russian).
2.2. V. I. Agoshkov, in: *Differential and Integrodifferential Equations*, (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1977): 44-72 (in Russian).
2.3. V. I. Agoshkov, *Some Questions of the Theory and Approximate Solution of Particle Transport Problems* (Moscow: Otdel Vychisl. Matem. AN SSSR, 1984) (in Russian).
2.4. V. S. Vladimirov, *Zh. Vychisl. Mat. i Mat. Fiz.* 8, No. 4: 842 (1968) (in Russian).
2.5. V. S. Vladimirov, *Vychisl. Matem.* No. 3: 3-33 (1958) (in Russian).
2.6. T. A. Germogenova, *Dokl. Akad. Nauk SSSR* 187: 5 (1969) (in Russian).
2.7. G. I. Marchuk and V. I. Agoshkov, *Dokl. Akad. Nauk SSSR* 232, No. 6: 1253-1256 (1977) (in Russian).
2.8. H. G. Kaper and R. B. Kellogg, *SIAM J. Appl. Math.* 32, No. 1: 201-214 (1977).
2.9. V. P. Shutyaev, in: *Numerical Methods and Statistical Modelling in Transport Theory* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980): 81-88 (in Russian).
2.10. V. I. Agoshkov, in: *Partial Differential Equations* (Novosibirsk: Nauka, 1980): 105-107 (in Russian).
2.11. V. I. Agoshkov, *Functions with Limited Variations and the Smoothness of Solutions of the Transport Equation* (Moscow: Otdel Vychisl. Mat. AN SSSR, Preprint No. 62, 1980) (in Russian).
2.12. V. I. Agoshkov, *Proceedings of the International Mathematical Congress* (Warsaw, 1983).
2.13. G. I. Marchuk and V. I. Agoshkov, *Introduction to Projection-Grid Methods* (Moscow: Nauka, 1981) (in Russian).
2.14. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd revised and enlarged ed., (Moscow: Nauka, 1980) (in Russian).
2.15. V. I. Agoshkov and V. P. Shutyaev, in: *Numerical Methods and Statistical Modelling in Transport Theory* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980): 19-34 (in Russian).

2.16. G. Strang and G. J. Fix, *An Analysis of the Finite Element Method* (Englewood Cliffs: Prentice-Hall, 1973).
2.17. L. A. Oganesyan and L. A. Rukhovets, *Variational Difference Methods for Solving Elliptic Equations* (Yerevan: Akad. Nauk. Armyan. SSR, 1979) (in Russian).
2.18. G. I. Marchuk and V. V. Shaidurov, *Increasing the Accuracy of Solutions of Difference Schemes* (Moscow: Nauka, 1979) (in Russian).
2.19. O. V. Besov, V. P. Ilyin and S. M. Nikol'sky, *Integral Representations and Embedding Theorems* (Moscow: Nauka, 1977) (in Russian).
2.20. S. M. Nikol'sky, *Approximation of Functions of Several Variables and Embedding Theorems* (Moscow: Nauka. 1977) (in Russian).
2.21. I. Babuška, *Approximation by Hill Functions* (Inst. for Fluid Dynam. Appl. Math. Univ. of Maryland, Tech. Note BN-648, 1970).
2.22. G. I. Marchuk and V. I. Lebedev, *Numerical Methods in Neutron Transport Theory* (Moscow: Atomizdat, 1971) (in Russian).

3 Variational Forms of Problems in Transport Theory

V. I. Agoshkov and V. V. Smelov

New results on the method of spherical harmonics and applications of adjoint equations and conjugate functions in transport theory are presented. These results are used in obtaining variational forms of problems in transport theory. First, variational formulations of problems for the method of spherical harmonics are considered. It is then shown how using adjoint equations we can obtain variational statements directly for the kinetic equation.

3.1. INTRODUCTION

A great number of methods of approximate solution of problems for the Boltzmann transport equation and approaches to the study of these problems have been developed in transport theory. The method of spherical harmonics (MSH) and the algorithms of the theory of conjugate functions are the most known. These algorithms have been intensively studied by Marchuk [3.1]-[3.4] and his pupils [3.5]-[3.7] and by many other authors.

In 1958 in [3.1] a vector-matrix representation of typical reactor problems in an arbitrary P_N approximation of the method of spherical harmonics was formulated in detail. The representation, which formally "copies" the classical diffusion equation, possesses many of the latter's properties. In particular, it allows the same finite-difference notation as does the diffusion equation. As a result a problem in the P_N approximation can be solved by the standard sweep algorithm (in a matrix-vector variant).

The next, very important, stage in the development of the MSH was the removal of the causes of "incorrectness" of even P_{2N} approximations, that were current prior to the 1970's. The incorrectness was due, in the first place, to the wrong ideas about the principles of imposing conditions of integration on the internal boundaries of a domain where physical parameters of a material medium underwent discontinuities. Investigations of Rumyantsev [3.3] and Marchuk [3.2] identified and removed the causes of the "inferiority" of P_{2N} approximations and as a result even approximations were "rehabilitated". All this first of all stimulated interest in the simplest P_2 approximation which substantially improves the diffusion theory [3.2].

Current developments in the MSH are reflected in [3.8] where additional attention is given to the construction of difference analogues of enhanced accuracy and to a thorougher study of the factorization method as a tool of solving finite-difference problems.

Among another group of algorithms widely used in reactor design theory are algorithms based on the use of adjoint equations and conjugate functions. The development of these algorithms is to a great extent due to [3.9] in which for the transport equation a theory of perturbations with respect to the functionals of a problem's solution has been developed. The approaches developed in that paper have substantially influenced the formation of the principles of numerical solution (using the theory of adjoint equations and conjugate functions) of a number of important problems of mathematical physics, the theory of neuclear reactors, meteorology and astrophysics [3.8], [3.10], [3.11].

This paper presents new results on the method of spherical harmonics and on applications of adjoint equations and conjugate functions in transport theory. The results concern obtaining variational forms of problems in transport theory. The paper first discusses variational formulations of problems in the method of spherical harmonics. It is then shown how using adjoint equations one can derive variational statements of problems directly for the kinetic equation.

The results presented in this paper have been obtained with the participation of, or under the influence of the approaches that have been developed for a number of years by G. I. Marchuk and these authors express their profound gratitude to him for the support of their work.

3.2. VARIATIONAL FORM OF THE METHOD OF SPHERICAL HARMONICS

3.2.1. The method of spherical harmonics. The linear Boltzmann equation describing steady-state one-velocity processes of particle transport in a medium can be represented as

$$L\varphi \equiv \Omega \nabla \varphi + \sigma(\mathbf{x})\,\varphi - \sigma_s(\mathbf{x}) \int_{\Omega} g(\mathbf{x},\ \Omega\Omega')\,\varphi(\mathbf{x},\ \Omega')\,d\Omega' = f(\mathbf{x},\ \Omega),$$

where $\varphi(\mathbf{x},\ \Omega)$ is the desired density of a particle flux, $\Omega \in \Omega$, (Ω is a set of unit vectors in R_3) $\mathbf{x} \in D$, $D \subset R_3$ is a spatial domain, $\Omega\Omega' = \mu$ is a scalar product of vectors Ω and Ω'; $\sigma(\mathbf{x}) > 0$, $\sigma_s(\mathbf{x}) \geqslant 0$, $g(\mathbf{x},\ \mu) > 0$, $f(\mathbf{x},\ \Omega)$ are given functions ($\sigma(\mathbf{x}) \geqslant \sigma_s(\mathbf{x})$,

$$\int_{-1}^{1} g(\mathbf{x},\ \mu)\,d\mu = 1/(2\pi).$$

The method of spherical harmonics (MSH) is the following algorithm (P_N approximation):

$$\text{(a)}\quad \Phi(\mathbf{x}, \Omega) = \sum_{k=0}^{N} \sum_{m=-k}^{k} \varphi_{km}(\mathbf{x})\, Y_k^{(m)}(\Omega), \tag{3.1}$$

$$\text{(b)}\quad \int_{\Omega} (L\Phi - f)\, \overline{Y_k^{(m)}}(\Omega)\, d\Omega = 0, \quad k = \overline{0, N}, \quad m = -\overline{k, k}, \tag{3.2}$$

where $\varphi_{km}(\mathbf{x})$ are desired coefficients and $\{Y_k^{(m)}(\Omega)\}$ is an orthonormal system of spherical functions (complex-valued in the general case*).

In 1961 V. S. Vladimirov formulated for the P_N approximation (with N odd) homogeneous boundary conditions corresponding to a zero irradiation of a region from the outside which ensure that the P_N approximation is symmetrizable when the scattering indicatrix $g(\mathbf{x}, \Omega\Omega')$ and the function of sources $f(\mathbf{x}, \Omega)$ are even with respect to Ω [3.12].

These results are developed below along the following lines: (i) symmetrization of an odd P_{2N+1} approximation under the Vladimirov boundary conditions (both homogeneous and inhomogeneous) is carried out without any constraints on the evenness of g and f, (ii) for the most general form of P_{2N+1} approximation conditions are found imposed by the symmetry or periodicity in the "input parameters" and under these conditions (again without the constraints mentioned above on the evenness with respect to Ω) the P_{2N+1} approximation is shown to be symmetrizable.

3.2.2. Integration conditions and the principle of obtaining boundary conditions in the MSH. 1. In what follows the odd P_{2N+1} approximation is mainly considered:

$$\Phi(\mathbf{x}, \Omega) = \sum_{k=0}^{2N+1} \sum_{m=-k}^{k} \varphi_{km}(\mathbf{x})\, Y_k^{(m)}(\Omega). \tag{3.3}$$

The even P_{2N} approximation is only briefly described at the end of this section.

The parity properties of spherical functions ($Y_k^{(m)}(-\Omega) = (-1)^k Y_k^{(m)}(\Omega)$) allow $\Phi(\mathbf{x}, \Omega)$ in (3.3) to be represented in a natural way as $\Phi(\mathbf{x}, \Omega) = \varphi^0(\mathbf{x}, \Omega) + \varphi^1(\mathbf{x}, \Omega)$, where $\varphi^i(\mathbf{x}, \Omega) = (-1)^i \varphi^i(\mathbf{x}, -\Omega)$, $i = 0, 1$. In particular,

*In the case of complex spherical functions coefficients φ_{km} in representations (3.1), (3.3), and (3.4) are complex, while the function $\Phi(\mathbf{x}, \Omega)$ is real.

$$\varphi^0(\mathbf{x}, \Omega) = \sum_{k=0}^{N} \sum_{m=-2k}^{2k} \varphi_{2k,\,m}(\mathbf{x})\, Y_{2k}^{(m)}(\Omega). \tag{3.4}$$

Let S be an arbitrary internal surface in D of class $C^{(1)}$ and let S^+ and S^- be its two sides. A jump of any function $\alpha(\mathbf{x}, \Omega)$ on S will be denoted by $[\alpha(\mathbf{x}, \Omega)]_S \equiv \alpha(\mathbf{x}_{S+}, \Omega) - \alpha(\mathbf{x}_{S-}, \Omega)$.

2. After these preliminaries it is possible to formulate the following result relating to integration conditions of Rumyantsev [3.3]:

Theorem 1. *In the case of the odd P_{2N+1} approximation the following three representations of integration conditions are equivalent:*

$$\int_{\Omega} \Omega\mathbf{n}\,[\Phi(\mathbf{x}, \Omega)]_S Y_k^{(m)}(\Omega)\, d\Omega = 0, \quad k = \overline{0,\ 2N+1},\ m = -\overline{k,\ k}, \tag{3.5}$$

$$\begin{cases} [\varphi^0(\mathbf{x}, \Omega)]_S = 0 \\ \displaystyle\int_{\Omega} \Omega\mathbf{n}\,[\varphi^1(\mathbf{x}, \Omega)]_S Y_{2k}^{(m)}(\Omega)\, d\Omega = 0, \quad k = \overline{0,\ N},\ m = -\overline{2k,\ 2k}, \end{cases} \tag{3.6}$$

$$\begin{cases} \displaystyle\int_{\Omega\mathbf{n}>0} \Omega\mathbf{n}\,[\Phi(\mathbf{x}, \Omega)]_S Y_{2k}^{(m)}(\Omega)\, d\Omega = 0, \\ \displaystyle\int_{\Omega\mathbf{n}<0} \Omega\mathbf{n}\,[\Phi(\mathbf{x}, \Omega)]_S Y_{2k}^{(m)}(\Omega)\, d\Omega = 0 \end{cases} \quad k = \overline{0,\ N},\ m = -\overline{2k,\ 2k}, \tag{3.7}$$

where $\mathbf{n}$ *is the unit normal to* S.

For the *proof* of the theorem see [3.13].

Of the three equivalent variants of integration conditions (3.5) to (3.7) the third has a structure in which the flow of a substance over the boundary S in the direction of the normal ($\Omega\mathbf{n} > 0$) and in the opposite direction ($\Omega\mathbf{n} < 0$) is separately taken into account. This fact allows us to "extract" directly from (3.7) the boundary conditions on the external (nonconcave) boundary ∂D in the P_{2N+1} approximation. Indeed if at points $\mathbf{x}_S \in \partial D$ the flux density $\varphi_{(\Gamma)}(\mathbf{x}_S, \Omega)$ of a substance irradiating a region D from the outside ($\varphi_{(\Gamma)}(\mathbf{x}_S, \Omega)$ is considered as $\Phi(\mathbf{x}_{S+}, \Omega)$) is known, then the second complex of conditions in (3.7) will be the physically justified system of boundary conditions:

$$\int_{\Omega\mathbf{n}<0} \Omega\mathbf{n}\,[\Phi_*(\mathbf{x}_S, \Omega) - \varphi_{(\Gamma)}(\mathbf{x}_S, \Omega)]\, Y_{2k}^{(m)}(\Omega)\, d\Omega = 0, \tag{3.8}$$

$$k = \overline{0,\ N}, \quad m = -\overline{2k,\ 2k},$$

where $\Phi(\mathbf{x}_S, \Omega)$ is considered as $\Phi(\mathbf{x}_{S-}, \Omega)$. In the general form represented by equations (3.8) the boundary conditions were for the first time obtained and substantiated by Vladimirov [3.12]. The theorem thus establishes a logical connection between the Vladimirov conditions (3.8) and the Rumyantsev conditions (3.5).

3. Since for the cases of symmetry ("mirror reflection") and periodicity the symmetry plane of a problem or the boundary of an element of periodicity are internal surfaces of a medium, the boundary conditions for these cases in the P_{2N+1} approximation can be obtained using the above theorem [3.6].

Consider in detail the first of these cases, i.e. the case where S is a symmetry plane. To begin with, we introduce an "auxiliary" orthogonal basis for even spherical functions $\Pi_{2k}^{(m)}(\Omega)$ satisfying the conditions $\Pi_{2k}^{(m)}(\Omega) = \varepsilon_{km}\Pi_{2k}^{(m)}(\Omega')$, where $\Omega' = \Omega - 2\,(\Omega\mathbf{n})\,\mathbf{n}$ is the mirror reflection of the vector Ω and ε_{km} is either 1 or -1. As $\Pi_{2k}^{(m)}(\Omega)$ one may take, for example, the spherical functions

$$\Pi_{2k}^{(m)}(\Omega) = P_{2k}^{(|m|)}(\mu)\,T_m(\psi),\quad T_m(\psi) = \begin{cases} \sin m\psi, & m > 0, \\ \cos m\psi, & m \leqslant 0, \end{cases} \tag{3.9}$$

where (for $r \in S$) the reference axes of the angles ψ and $\theta = \cos^{-1}\mu$ are in the symmetry plane S. In the above example $\varepsilon_{km} = -\operatorname{sign} m$ $(m \neq 0)$, $\varepsilon_{k0} = 1$. It is easy to see that in any variant of auxiliary basis $\{\Pi_{2k}^{(m)}(\Omega)\}$ among the numbers $\{\varepsilon_{km}\}_{m=-2k}^{2k}$ there will be $2k+1$ positive units and $2k$ negative units.

From the even basic $\{\mathbf{Y}_{2k}^{(m)}(\Omega)\}$ and the auxiliary $\{\Pi_{2k}^{(m)}(\Omega)\}$ systems of spherical functions, as well as from the coefficients $\{\varphi_{2k,m}(\mathbf{x}_S)\}$ $(\mathbf{x}_S \in S)$ we form vector functions $\mathbf{Y}_k(\Omega)$, $\mathbf{\Pi}_k(\Omega)$, $\varphi_k(\mathbf{x}_S)$ as follows:

$$\mathbf{Y}_k(\Omega) = (Y_{2k}^{(-2k)}(\Omega),\ Y_{2k}^{(-2k+1)}(\Omega),\ \ldots,\ Y_{2k}^{(2k)}(\Omega))^T,\ \text{etc.},$$

where T is transposition.

Vector functions $\mathbf{Y}_k(\Omega)$ and $\mathbf{\Pi}_k(\Omega)$ are linearly expressed in terms of each other: $\mathbf{\Pi}_k(\Omega) = B_k\mathbf{Y}_k(\Omega)$, where B_k is a nonsingular Ω-independent matrix. Finally, denote by E_k a diagonal matrix: $E_k = \operatorname{diag}(\varepsilon_{k,-2k},\ \varepsilon_{k,-2k+1},\ \ldots,\ \varepsilon_{k,2k})$.

Proceeding now from the first equation $[\varphi^0(\mathbf{x}, \Omega)]_S = 0$ in representation (3.6) of the theorem, it is easy to obtain in the symmetry plane S the following complex of boundary conditions [3.4]:

$$Q_kB_k\varphi_k(\mathbf{x}_S) = 0,\quad k = \overline{0, N}, \tag{3.10}$$

where $Q_k = (E - E_k)/2$ is a diagonal matrix with $2k+1$ zeros and $2k$ unities on the diagonal. (Here E is a unit matrix of order $4k+1$.) System (3.10) thus contains only $2k$ nontrivial conditions. The second complex of conditions in (3.6) is converted to the form [3.4]

$$(1+\varepsilon_{km})\int_\Omega \Omega\mathbf{n}\varphi^1(\mathbf{x}_S, \Omega)\Pi_{2k}^{(m)}(\Omega)\,d\Omega = 0,\ k = \overline{0, N},\quad m = -\overline{2k, 2k}, \tag{3.11}$$

where the number of nontrivial $(\varepsilon_{km} = 1)$ relations for every k equals $2k+1$.

Remark 1. If the integration conditions (3.6) are formally (without any additional transformations) transformed in the symmetry plane to boundary conditions, then there will occur some extra (linearly dependent) ones among them. Algorithm (3.10), (3.11) allows constructing a system of boundary conditions by singling out a minimum equivalent subsystem in the original system of conditions.

In conclusion it is in order to mention a particular variant of writing "mirror reflection" conditions when P_{2N+1} approximation is formed using spherical functions of the form (3.9) and the reference axes of the angles ψ and $\theta = \cos^{-1}\mu$ for $r \in S$ are in S. In this case the symmetry conditions reduce to the equations

$$\varphi_{k,m}(\mathbf{x}_S) = 0, \quad k = \overline{1,\ 2N+1}, \quad m = \overline{1,\ k}$$

from which the following consequence can be derived:

$$\left.\begin{aligned} \varphi_{2k,m}(\mathbf{x}_S) &= 0, \quad m > 0, \\ \partial\varphi_{2k,m}/\partial n\,|\mathbf{x}_S &= 0, \quad m \leqslant 0, \end{aligned}\right\} \quad k \leqslant \overline{1,\ N}.$$

3.2.3. Variational form of the odd P_{2N+1} approximation. 1. With respect to real functions of the form (3.3) let us agree to write $\Phi \in \mathscr{H}(D)$ if $\forall\, \varphi_{km} \in W_2^1(D)$. Accordingly, for real functions of the form (3.4) we shall use the notation

$$\varphi^0 \in \mathscr{H}^0(D) \quad \text{if} \quad \forall \varphi_{2k,m} \in W_2^1(D),$$
$$\varphi^0 \in \mathscr{L}_2(D) \quad \text{if} \quad \forall \varphi_{2k,m} \in L_2(D).$$

Finally, let us agree to write

$$\varphi^1 \in \mathscr{H}^1(D) \text{ if } \forall \varphi_{2k+1,m} \in W_2^1(D).$$

2. We now take an arbitrary function $\Phi \in \mathscr{H}(D)$ and using it obtain the right-hand sides in the following expressions (reflecting an expanded notation of problem (3.2) (with $k = \overline{0,\ 2N+1}$) and (3.8) (see below)):

$$\int_\Omega \overline{Y_{2k}^{(m)}}\,\Omega\nabla\varphi^1\,d\Omega + \sigma_{2k}\varphi_{2k,m} = f_{2k,m}, \quad m = -\overline{2k,\ 2k}, \tag{3.12}$$

$$\int_\Omega \overline{Y_{2k+1}^{(m)}}\,\Omega\nabla\varphi^0\,d\Omega + \sigma_{2k+1}\varphi_{2k+1,m} = f_{2k+1,m}, \quad m = -\overline{2k-1,\ 2k+1}, \tag{3.13}$$

$$\int_{\Omega n<0} \Omega n\Phi(\mathbf{x}_S,\ \Omega)\,\overline{Y_{2k}^{(m)}}\,d\Omega = \varphi_{2,m}^{(-)}(\mathbf{x}), \quad m = -\overline{2k,\ 2k}, \tag{3.14}$$

where $k = \overline{0,\ N}$, $\mathbf{x}_S \in \partial D$, and $\sigma_0(\mathbf{x}) \geqslant 0$, $\sigma_\nu(\mathbf{x}) \geqslant \varkappa > 0$ $(\nu > 0)$ are some functions in $L_\infty(D)$.

By the definition of $\mathscr{H}(D)$ the left-hand sides of system (3.12) to (3.14) are meaningful. It is also clear that in (3.12) to (3.14) for all k and m

$$f_{2k,m},\ f_{2k+1,m} \in L_2(D), \quad \varphi^{(-1)}_{2k,m} \in L_2(\partial D).$$

We write system (3.12) to (3.14) in operator form:

$$A\Phi = \mu, \tag{3.15}$$

where μ is a vector function of the right-hand sides of (3.12) to (3.14). If Φ (the inverse image) runs over the entire space $\mathscr{H}(D)$, then the collection of all elements μ (images) which is defined by equation (3.15) forms some set $\mathscr{M}$, i.e. $A\mathscr{H} = \mathscr{M}$. $\Phi \Rightarrow \mu$ is a unique correspondence, uniqueness in the reverse direction following from the subsequent results.

3. We now turn to problem (3.2) (with $k = \overline{0,\ 2N+1}$) and (3.8). Considering the equations [3.6]

$$\int_\Omega Y^{(m)}_{2k}\,\Omega\,\nabla\varphi^0\,d\Omega = \int_\Omega Y^{(m)}_{2k+1}\Omega\nabla\varphi^1\,d\Omega = 0$$

its expanded notation assumes the form (3.12) to (3.14) with

$$\begin{aligned} f_{\nu m}(\mathbf{x}) &= \int_\Omega f(\mathbf{x},\ \Omega)\,\overline{Y^{(m)}_\nu}(\Omega)\,d\Omega \\ \varphi^{(-)}_{\nu m}(\mathbf{x}_S) &= \int_{\Omega\mathbf{n}<0} \Omega\mathbf{n}\varphi_{(\Gamma)}(\mathbf{x}_S,\ \Omega)\,\overline{Y^{(m)}_\nu}(\Omega)\,d\Omega. \end{aligned} \tag{3.16}$$

It is essential that at this stage (up to stage 4) it shall be assumed that the vector function μ calculated using (3.16) is a member of the set $\mathscr{M}$. This ensures the existence of (at least one!) member $\Phi \in \mathscr{H}(D)$ which is a solution of problem (3.12) to (3.14).

Since $\sigma_{2k+1}(x) \neq 0$, using (3.13) each of the functions $\varphi_{2k+1,m}(\mathbf{x}) \in W_2^1(D)$ can be explicitly expressed and substituted into the other equations of the system. As a result problem (3.12) to (3.14) becomes

$$\int_\Omega \overline{Y^{(m)}_{2k}}\Omega\nabla\varphi^1\,d\Omega + \sigma_{2k}\varphi_{2k,m} = f_{2k,m}, \tag{3.17}$$

$$\int_{\Omega\mathbf{n}<0} \Omega\mathbf{n}\,[\varphi^0(\mathbf{x}_S,\ \Omega) + \varphi^1(\mathbf{x}_S,\ \Omega)]\,\overline{Y^{(m)}_{2k}}\,d\Omega = \varphi^{(-)}_{2k,m}(\mathbf{x}_S),$$

$$k = \overline{0,\ N}, \quad m = \overline{-2k,\ 2k}, \tag{3.18}$$

where

$$\varphi^1(\mathbf{x},\ \Omega) = -\sum_{k=0}^{N}\sum_{m=-2k-1}^{2k+1}\left(\int_\Omega \overline{Y^{(m)}_{2k+1}}\Omega\nabla\varphi^0\,d\Omega - f_{2k+1,m}\right)Y^{(m)}_{2k+1}/\sigma_{2k+1}. \tag{3.19}$$

Remark 2. Since the right-hand side of (3.19) is acted on by the operator ∇ (see [3.17]), it is important to stress that this operator shall act on the combinations

$$\left(\int_{\Omega} \overline{Y_{2k+1}^{(m)}} \Omega \nabla \varphi^0 \, d\Omega - f_{2k+1,\, m}\right)/\sigma_{2k+1} \equiv -\varphi_{2k+1,\, m} \in W_2^1(D)$$

only in the large rather than separately on each of the terms. The expression in parentheses in (3.19) can be separated only after the effect of ∇ on φ^1 has been removed (see below).

We transform the set of functions $\mathscr{L}_2(D)$ introduced above into a Hilbert space with scalar product

$$(\varphi^0, \psi^0) = \int_{D\times\Omega} \varphi^0(\mathbf{x}, \Omega)\, \psi^0(\mathbf{x}, \Omega)\, d\mathbf{x}\, d\Omega$$

$$= \sum_{k=0}^{N} \sum_{m=-2k}^{2k} \int_D \varphi_{2k,\, m}(\mathbf{x}) \bar{\psi}_{2k,\, m}(\mathbf{x})\, d\mathbf{x}. \quad (3.20)$$

Using the second representation for the scalar product in (3.20) perform a scalar multiplication of (3.17) by an arbitrary element $v \in \mathscr{H}^0(D)$. We reprent the combination $v(\Omega\nabla\varphi^1)$ in the first integrand by the right-hand side of the identity $v(\Omega\nabla\varphi^1) \equiv \nabla(\Omega v\varphi^1) - \varphi^1(\Omega\nabla v)$ valid by virtue of the inclusions $\varphi^1 \in \mathscr{H}^1(D)$, $v \in \mathscr{H}^0(D)$. Using next the divergence theorem we get as a result

$$K_1[\varphi^0, v] + \tilde{K}_2[\varphi^0, v] + \tilde{K}_3[\varphi^0, v] - (f^0, v) = 0, \quad (3.21)$$

where f^0 is the analogue of φ^0 and

$$K_1[\varphi^0, v] = \sum_{k=0}^{N} \sum_{m=-2k}^{2k} \sigma_{2k} \varphi_{2k,\, m} \bar{v}_{2k,\, m},$$

$$\tilde{K}_2[\varphi^0, v] = \int_{\partial D} ds \int_{\Omega} \Omega \mathbf{n} \varphi^1(\mathbf{x}, \Omega)\, v(\mathbf{x}, \Omega)\, d\Omega, \quad (3.22)$$

$$\tilde{K}_3[\varphi^0, v] = -\int_{D\times\Omega} \varphi^1(\mathbf{x}, \Omega)\, [\Omega\nabla v(\mathbf{x}, \Omega)]\, d\mathbf{x}\, d\Omega$$

($\tilde{K}_2$ and $\tilde{K}_3$ are functionals of φ^0 since it is assumed that φ^1 are eliminated from (3.12) by means of (3.19)).

Note that in (3.21) φ^1 is no longer acted upon by ∇ (see (3.22)), its action "switched over" onto v (see the second equation in (3.22)). Now the last sentence of Remark 2 becomes valid.

The integrals in (3.22) are subjected to identity transformations [3.6] based on (3.18) and (3.19) and the properties of evenness of φ^0

and φ^1. As a result equation (3.21) becomes $K[\varphi^0, v] = F(v)$, where

$$K[\varphi^0, v] = K_1[\varphi^0, v] + K_2[\varphi^0, v] + K_3[\varphi^0, v], \tag{3.23}$$

$$K_2[\varphi^0, v] = 2\int_{\partial D} ds \int_{\Omega \mathbf{n}>0} \Omega \mathbf{n} \varphi^0(\mathbf{x}, \Omega) v(\mathbf{x}, \Omega)\, d\Omega,$$

$$K_3[\varphi^0, v] = \sum_{k=0}^{N} \sum_{m=-2k-1}^{2k+1} \int_D (1/\sigma_{2k+1}) \left(\int_\Omega \overline{Y^{(m)}_{2k+1}} \Omega \nabla \varphi^0 \, d\Omega \right) \times \left(\int_\Omega Y^{(m)}_{2k+1} \Omega \nabla v \, d\Omega \right) dx, \tag{3.24}$$

$$F(v) = (f^0, v) - 2\int_{D\partial} ds \int_{\Omega \mathbf{n}<0} \Omega \mathbf{n} \varphi_{(\Gamma)}(\mathbf{x}, \Omega) v(\mathbf{x}, \Omega)\, d\Omega + \sum_{k=0}^{N} \sum_{m=-2k-1}^{2k+1} \int_D (f_{2k+1, m}/\sigma_{2k+1})\, dx \int_\Omega Y^{(m)}_{2k+1} \Omega \nabla v \, d\Omega. \tag{3.25}$$

$K[u, v]$ with $u, v \in \mathscr{H}^0(D)$ arbitrary is a symmetric bilinear form which for $u = v$ leads to $K[v, v] \geqslant 0$. It is shown in [3.6] that the last inequality can be strengthened: $K[v, v] \geqslant \gamma (v, v)$, where $\gamma > 0$ is independent of the approximation number N. By virtue of the last estimate, it follows from the general theory [3.14] that $\varphi^0(\mathbf{x}, \Omega)$ is a unique solution of the variational problem

$$J(v) \equiv K[v, v] - 2F(v) = \min, \quad v \in \mathscr{H}^0(D). \tag{3.26}$$

The function $\varphi^1(\mathbf{x}, \Omega)$ is then uniquely characterized using equation (3.19).

So if the right-hand side μ of problem (3.15) is such that $\mu \in \mathscr{M}$, then that problem (or equivalently problem (3.12) to (3.14)) is uniquely solvable, with $\varphi^0 \in \mathscr{H}^0(D)$, $\varphi^1 \in \mathscr{H}^1(D)$.

4. Suppose now that the functions $f_{\nu m}(\mathbf{x})$ and $\varphi^{(-)}_{\nu m}(\mathbf{x}_S)$ calculated using (3.16) are such that $\mu \in \mathscr{M}$. In this (unfavourable) situation there is no solution of problem (3.17) to (3.19) in the straightforward (classical) sense. In such cases it is customary to go over to the concept of generalized solution by which we shall mean the solution of the variational problem (3.26). For such a solution to exist and be unique, besides the positive definiteness of $K(v, v]$ it is required that additional prerequisites should be met [3.14]. In our problem they are satisfied:

(i) $\mathscr{H}^0(D)$ is a complete space in the energy norm $\| v \|^2 = K[v, v]$;

(ii) $\mathscr{H}^0(D)$ is everywhere dense in $\mathscr{L}_2(D)$ in the norm defined in terms of the scalar product (3.20);

(iii) the additive functional (3.25) is bounded in the energy norm: $F^2(v) \leqslant \text{const} \cdot K[v, v]$.

Remark 3. Given the generalized solution of the problem, the function $\varphi^1(\mathbf{x}, \Omega)$ defined by formula (3.19) shall no longer be in the space $\mathscr{H}^1(D)$.

5. Similarly justified is the variational formulation of the problem in the case of symmetry or periodicity conditions [3.6]. In particular, under symmetry conditions the problem is reduced to minimizing the functional

$$J(v) \equiv K[v, v] - 2F(v), \quad v \in \mathscr{H}_c^0(D), \tag{3.27}$$

where $K[v, v]$ is formally derived from (3.23) (with $K_2 = 0$) and $F(v)$ is derived from formula (3.25) (with the surface integral zero). The space $\mathscr{H}_c^0(D)$ in (3.27) is defined as follows:

$$\mathscr{H}_c^0(D) = \{v(\mathbf{x}, \Omega): v \in \mathscr{H}^0(D), \quad Q_k B_k v_k(\mathbf{x}_S) = 0, \quad k = \overline{0, N}\},$$

where Q_k and B_k are matrices in (3.10) and $v_k(\mathbf{x}_S)$ is a vector with components $v_{2k,m}(\mathbf{x}_S)$, $m = -2k, 2k$. Minimization of functional (3.27) is carried out in the range of $\mathscr{H}_c^0(D)$ and hence requires that $v(\mathbf{x}, \Omega)$ should obey the boundary conditions $Q_k B_k v_k(\mathbf{x}_S) = 0$, $k = \overline{0, N}$. These are by definition [3.14] *principal* conditions while conditions (3.11) are *natural*.

In the case of the periodic variant the variational functional completely coincides in construction with (3.27), except that the trial functions v are chosen from another space, $\mathscr{H}_c^0(D)$. That space is formed from functions $v(\mathbf{x}, \Omega) \in \mathscr{H}^0(D)$ satisfying the following (principal) condition: $v(\mathbf{x}_S, \Omega) = v(\mathbf{x}_S', \Omega)$, where $\mathbf{x}_S$ and $\mathbf{x}_S'$ are mutually congruent points of the boundary ∂D of a periodicity element.

Remark 4. If on different sections of ∂D the solution of a problem satisfies different boundary conditions of those listed above (for example, Vladimirov conditions on one section and symmetry conditions on another), then the entire variational "technique" and its justification can be carried over without difficulty to this combined case as well: in formulas (3.24) and (3.25) one must understand by ∂D only that part of the boundary where a Vladimirov condition is given (the trial function is not required to obey boundary conditions there); at the other points of the boundary the trial functions must be subjected to the corresponding principal boundary conditions.

In conclusion we note that (unlike Vladimirov conditions) the symmetry or periodicity conditions cannot by themselves guarantee even the existence of a solution of the stationary problem and hence the positive definiteness of the quadratic form $K[v, v]$. In this situation the inequality $K[v, v] \geqslant \gamma(v, v)$, $(\gamma > 0)$, and therefore a unique solvability of the problem are possible only under some additional sufficient conditions. Specifically, it suffices that any of

the following two conditions are met:

(i) $\sigma(\mathbf{x}) - \sigma_S(\mathbf{x}) \geqslant \alpha > 0$, $x \in D_0$, $D_0 \subseteq D$;

(ii) ∂D contains a section $S \subset \partial D$ where the Vladimirov condition (3.8) is given.

6. The above results for the odd P_{2N+1} approximation can be obtained for the even P_{2N} approximation as well. This is based on the principle of duality which, for example, for the P_{2N} approximation too leaves the theorem and the boundary conditions (3.8) of Sec. 3.2.2 valid if in (3.6) to (3.8) the indices $2k$ are replaced by $2k + 1$ and in (3.6) in addition the superscripts 0 and 1 are interchanged. For the even P_{2N} approximation all studies and justifications were carried out in [3.15], [3.16].

3.3. THE VARIATIONAL FORM OF PROBLEMS FOR THE TIME-DEPENDENT KINETIC EQUATION

It is shown here how to use the adjoint equation to symmetrize the problem for the time-dependent transport equation and to reduce the problem (with or without its conjugate) to a problem with a self-adjoint operator, which in turn reduces to the problem of finding a minimum of the quadratic functional. The exposition is made using an example of the time-dependent one-velocity transport equation with isotropic scattering indicatrix. The procedure to be described can also be applied, however, to time-dependent (and time-independent) kinetic equations with energy dependence and anisotropic indicatrixes and to a number of other equations of mathematical physics.

3.3.1. Definitions and subsidiary statements. Suppose in what follows that D is a convex bounded domain with a piecewise smooth boundary. By t is understood a time variable taking values in the range $[0, T]$, $T < \infty$. By $\mathbf{n} = (n_1, n_2, n_3)$ we denote the unit vector of the outward normal to ∂D. If Ω is some fixed direction, then ∂D_- denotes those points of ∂D at which $(\Omega\mathbf{n}) \equiv \sum_{i=1}^{3} \Omega_i n_i \leqslant 0$ and ∂D_+ those points at which $(\Omega\mathbf{n}) > 0$. To within the null set we have $\partial D = \partial D_- \cup \partial D_+$.

In what follows the variable t will often be also denoted by x_0: $x_0 \equiv t$. Suppose now that $\mathbf{r}$ is a vector $\mathbf{r} = (x_0, x_1, x_2, x_3)$ assuming values in a four-dimensional half-space $(x_0 \equiv t \geqslant 0) \times R_3$ in which the axis Ox_0 is assumed to be orthogonal to the axes Ox_i, $i = 1, 2, 3$. The positive direction of Ox_0 is taken to coincide with the direction of increasing t. The unit vector in the positive direction of Ox_0 will be denoted by Ω_0, $|\Omega_0| = \Omega_0 = 1$.

Let $Q_T = (0, T) \times D$ be a four-dimensional cylinder with bounda-

ry $\partial Q_T = [(0, T) \times \partial D] \cup [(t = 0) \times D] \cup [(t = T) \times D]$. The unit vector of the outward normal to ∂Q_T is denoted by $\mathbf{N} = (n_0, n_1, n_2, n_3) = (n_0, \mathbf{n})$, where $n_0 = \pm\Omega_0$. Taking into account the orthogonality of Ox_0 to Ox_i, $i = 1, 2, 3$, we obtain a representation for $\mathbf{N}$ which is true almost everywhere on ∂Q_T:

$$\mathbf{N} = \begin{cases} (0, n_1, n_2, n_3) & t \in [0, T], x \in \partial D, \\ (n_0, 0, 0, 0), & t = (0, T), x \in D. \end{cases} \tag{3.28}$$

Let us also define a four-dimensional vector of the form

$$\mathbf{S} = (\Omega_0/c, \Omega_1, \Omega_2, \Omega_3) = (\Omega_0/c, \boldsymbol{\Omega}) \equiv (s_0, s_1, s_2, s_3), \tag{3.29}$$
$$|\mathbf{S}| = \sqrt{1/c^2 + 1},$$

where c is a positive constant. If now the direction of $\mathbf{S}$ is fixed, then $\partial Q_{T,-}$ denotes the set of points on ∂D_T for which $(\mathbf{S}\cdot\mathbf{N}) \leqslant 0$ and $\partial Q_{T,+}$ the set of points on ∂Q_T for which $(\mathbf{S}\cdot\mathbf{N}) > 0$. To within the null set (in a three-dimensional space) ∂Q_T can be represented as $\partial Q_T = \partial Q_{T,-} \cup \partial Q_{T,+}$, where $\partial Q_{T,-}$ and $\partial Q_{T,+}$ can also be defined by the relations

$$\partial Q_{T,-} = [(t = 0) \times D] \cup [(0, T) \times \partial D_-],$$
$$\partial Q_{T,+} = [(t = T) \times D] \cup [(0, T) \times \partial D_+] \tag{3.30}$$

or

$$\partial Q_{T'-} = \{\mathbf{r} \in \partial Q_T, (\mathbf{S}\cdot\mathbf{N}) \leqslant 0\},$$
$$\partial Q_{T,+} = \{\mathbf{r} \in \partial Q_T, (\mathbf{S}\cdot\mathbf{N}) > 0\}. \tag{3.31}$$

Suppose on all of the sets introduced above Lebesgue measures and the corresponding integration processes are defined.

We now introduce some real functional spaces obtained as completions of functions that are infinitely differentiable in $\overline{Q}_T = Q_T + \partial Q_T$ in the corresponding metrics.

$L_2 = L_2(Q_T \times \Omega)$ is a Hilbert space of square integrable functions $v = v(t, \mathbf{x}, \boldsymbol{\Omega})$ defined on $Q_T \times \Omega$:

$$(v, w) = \int\limits_{Q_T \times \Omega} vw\, dt\, d\mathbf{x}\, d\boldsymbol{\Omega}, \quad \|v\| = (v, v)^{1/2} < \infty.$$

$L_2(0, T; X)$ is a space of (classes of) functions v on $(0, T)$ which assume values in a Hibert space X:

$$(v, w)_{L_2(0, T; X)} = \int\limits_0^T (v, w)_X\, dt, \quad \|v\|_{L_2(0, T; X)} = \left(\int\limits_0^T \|v\|_X^2\, dt\right)^{1/2} < \infty.$$

The space H^1 is defined as follows:

$$H^{(1)} = \{v:\ (v,\ w)_{H(1)} = (v,\ w) + ((\mathbf{S}\cdot\nabla)\,v,\ (\mathbf{S}\cdot\nabla)\,w),$$

$$\| v \|_{H(1)} = (v,\ v)^{1/2}_{H(1)} < \infty\},$$

$$\nabla \equiv \left(\frac{\partial}{\partial x_0},\ \frac{\partial}{\partial x_1},\ \frac{\partial}{\partial x_2},\ \frac{\partial}{\partial x_3}\right) = \left(\frac{1}{c}\ \frac{\partial}{\partial t},\ \frac{\partial}{\partial x_1},\ \frac{\partial}{\partial x_2},\ \frac{\partial}{\partial x_3}\right).$$

The Hilbert spaces $L_2(\Omega \times \partial Q_{T,-})$, $L_2(\Omega \times \partial Q_{T,+})$, and $L_2(\Omega \times \partial Q_T)$ have, respectively, the norms and scalar products defined as follows:

$$(v,\ w)_{L_2(\Omega\times\partial Q_{T,\,-})} = \int\limits_{\Omega\times\partial Q_{T,\,-}} |(\mathbf{S}\cdot\mathbf{N})|\,vw\,dt\,\mathbf{x}\,d\Omega$$

$$= \int\limits_{\Omega\times D} (1/c)\,v\,(0,\ \mathbf{x},\ \Omega)\,w\,(0,\ \mathbf{x},\ \Omega)\,d\mathbf{x}\,d\Omega$$

$$+ \int\limits_0^T dt \int\limits_\Omega d\Omega \int\limits_{\partial D_-} |(\Omega\cdot\mathbf{n})|\,v\,(t,\ \mathbf{x},\ \Omega)\,w\,(t,\ \mathbf{x},\ \Omega)\,d\mathbf{x},$$

$$(v,\ w)_{L_2(\Omega\times\partial Q_{T,\,+})} = \int\limits_{\Omega\times\partial Q_{T,\,+}} |(\mathbf{S}\cdot\mathbf{N})|\,vw\,dt\,d\mathbf{x}\,d\Omega$$

$$= \int\limits_{\Omega\times D} (1/c)\,v\,(T,\ \mathbf{x},\ \Omega)\,w\,(T,\ \mathbf{x},\ \Omega)\,d\mathbf{x}\,d\Omega$$

$$+ \int\limits_0^T dt \int\limits_\Omega d\Omega \int\limits_{\partial D_+} |(\Omega\cdot\mathbf{n})|\,v\,(t,\ \mathbf{x},\ \Omega)\,w\,(t,\ \mathbf{x},\ \Omega)\,d\mathbf{x},$$

$$(v,\ w)_{L_2(\Omega\times\partial Q_T)} = (v,\ w)_{L_2(\Omega\times\partial Q_{T,\,-})} + (v,\ w)_{L_2(\Omega\times\partial Q_{T,\,+})},$$

$$\| v \|_{L_2(\Omega\times\partial Q_{T,\,-})} = (v,\ v)^{1/2}_{L_2(\Omega\times\partial Q_{T,\,-})} < \infty,$$

$$\| v \|_{L_2(\Omega\times\partial Q_{T,\,+})} = (v,\ v)^{1/2}_{L_2(\Omega\times\partial Q_{T,\,+})} < \infty,$$

$$\| v \|_{L_2(\Omega\times\partial Q_T)} = \{\| v \|^2_{L_2(\Omega\times\partial Q_{T,\,-})} + \| v \|^2_{L_2(\Omega\times\partial Q_{T,\,+})}\}^{1/2} < \infty.$$

We formulate a number of statements which justify the operations of integration (integration of the derivative $(\mathbf{S}\cdot\nabla)\,v$, integration by parts, etc.) to be used in what follows. For the proof of the statements see [3.7].

Lemma 3.1. *If* $\varphi, \psi \in H^{(1)}$ *and have on* ∂Q_T *traces belonging to the*

space $L_2(\Omega \times \partial Q_T)$, *then*

$$\int_{Q_T \times \Omega} (\mathbf{S}\cdot\nabla)\,\varphi(\mathbf{r},\ \Omega)\,d\mathbf{r}\,d\Omega = \int_{\Omega \times \partial Q_{T,+}} (\mathbf{S}\cdot\mathbf{N})\,\varphi\,d\mathbf{r}\,d\Omega + \int_{\Omega \times \partial Q_{T,-}} (\mathbf{S}\cdot\mathbf{N})\,\varphi\,d\mathbf{r}\,d\Omega, \quad (3.32)$$

$$\int_{Q_T \times \Omega} \psi\,(\mathbf{S}\cdot\nabla)\,\varphi\,d\mathbf{r}\,d\Omega = \int_{\partial Q_T \times \Omega} (\mathbf{S}\cdot\mathbf{N})\,\psi\varphi\,d\mathbf{r}\,d\Omega - \int_{Q_T \times \Omega} \varphi\,(\mathbf{S}\cdot\nabla)\,\psi\,d\mathbf{r}\,d\Omega, \quad (3.33)$$

where

$$\int_{\partial Q_T \times \Omega} (\mathbf{S}\cdot\mathbf{N})\,w\,d\mathbf{r}\,d\Omega = \int_{\Omega \times \partial Q_{T,+}} (\mathbf{S}\cdot\mathbf{N})\,w\,d\mathbf{r}\,d\Omega + \int_{\Omega \times \partial Q_{T,-}} (\mathbf{S}\cdot\mathbf{N})\,w\,d\,d\mathbf{r}\Omega,$$

$$\int_{\Omega \times \partial Q_{T,+}} (\mathbf{S}\cdot\mathbf{N})\,w\,d\mathbf{r}\,d\Omega = \int_{D \times \Omega} (1/c)\,w(T,\ \mathbf{x},\ \Omega)\,d\mathbf{x}\,d\Omega + \int_0^T dt \int_\Omega d\Omega \int_{\partial D_+} (\Omega\cdot\mathbf{n})\,w\,d\mathbf{x}, \quad (3.34)$$

$$\int_{\Omega \times \partial Q_{T,-}} (\mathbf{S}\cdot\mathbf{N})\,w\,d\mathbf{r}\,d\Omega = -\int_{D \times \Omega} (1/c)\,w(0,\ \mathbf{x},\ \Omega)\,d\mathbf{x}\,d\Omega + \int_0^T dt \int_\Omega d\Omega \int_{\partial D_-} (\Omega\cdot\mathbf{n})\,w\,d\mathbf{x}.$$

Lemma 3.2. *If* $\varphi \in H^{(1)}$ *assumes on* $\partial Q_{T,-}$ *(or on* $\partial Q_{T,+}$*) a boundary value* $g \in L_2(\Omega \times \delta Q_{T,-})$ *(or* $g \in L_2(\Omega \times \partial Q_{T,+})$*), then* v *has on* $\partial Q_{T,+}$ *(on* $\partial Q_{T,-}$*) a trace*

$$v\,|_{\partial Q_{T,+}} \in L_2(\Omega \times \partial Q_{T,+}) \quad (v\,|_{\partial Q_{T,-}} \in L_2(\Omega \times \partial Q_{T,-})).$$

Remark. In what follows, for the function $v\,|_{\partial Q_{T,+}}$, which has been called the trace of v on $\partial Q_{T,+}$, we shall write

$$\|\,v\,|_{\partial Q_{T,+}}\|_{L_2(\partial Q_{T,+} \times \Omega)} \equiv \|\,v\,\|_{L_2(\Omega \times \partial Q_{T,+})}.$$

3.3.2. Formulation of original problems and their transformation. Consider a time-dependent problem for a one-velocity kinetic

equation of the form

$$\frac{1}{c}\frac{\partial\varphi}{\partial t}+\sum_{i=1}^{3}\Omega_i\frac{\partial\varphi}{\partial x_i}+\sigma\varphi=\frac{\sigma_s}{4\pi}\int_{\Omega}\varphi\,d\Omega'+f$$

$$\varphi=\begin{cases}\varphi_{(\Gamma)} & \text{for } x\in\partial D,\quad (\Omega\cdot\mathbf{n})<0,\\ \varphi_{(0)} & \text{for } t=0,\end{cases} \tag{3.35}$$

where $f(t, \mathbf{x}, \Omega)\in L_2$, $\varphi_{(\Gamma)}(t, \mathbf{x}, \Omega)\in L_2(\Omega\times\partial D_-)$, $\varphi_{(0)}(\mathbf{x}, \Omega)\in L_2(D\times\Omega)$. It is assumed that for the coefficients $\sigma(t, \mathbf{x})$, $\sigma_s(t, \mathbf{x})$ the following relations hold:

$$0<\sigma_0\leqslant\sigma(t, \mathbf{x})\leqslant\sigma_1<\infty,\quad 0\leqslant\sigma_s(t, \mathbf{x})\leqslant\sigma_{s,1}<\infty,$$
$$0<\sigma_{c0}\leqslant\sigma_c(t, \mathbf{x})\equiv\sigma(t, \mathbf{x})-\sigma_s(t, \mathbf{x})\leqslant\sigma_{c1}<\infty^*$$

(σ_0, σ_1, σ_{s1}, σ_{c0}, σ_{c1} = const).

The constant c is assumed to be positive and it appears in the expression for the vector $\mathbf{S}$ given by equation (3.29). The problem conjugate to (3.35) is of the form

$$-\frac{1}{c}\frac{\partial\psi}{\partial t}-\sum_{i=1}^{3}\Omega_i\frac{\partial\psi}{\partial x_i}+\sigma\psi=\frac{\sigma_s}{4\pi}\int_{\Omega}\psi\,d\Omega'+g(t, \mathbf{x}, \Omega),$$

$$\psi=\psi_{(\Gamma)}(t, \mathbf{x}, \Omega),\quad \mathbf{x}\in\partial D,\quad (\Omega\cdot\mathbf{n})>0, \tag{3.36}$$

$$\psi(T, \mathbf{x}, \Omega)=\psi_{(1)}(\mathbf{x}, \Omega),$$

where it is assumed that $g\in L_2$, $\psi_{(\Gamma)}\in L_2(\Omega\times\partial D_+)$, $\psi_{(1)}\in L_2(D\times\Omega)$.

We define the following functions:

$$\gamma_{(-)}=\begin{cases}\varphi_{(\Gamma)} & \text{for } \mathbf{x}\in\partial D,\ (\mathbf{S}\cdot\mathbf{N})=(\Omega\cdot\mathbf{n})<0,\\ \varphi_{(0)} & \text{for } t=0\quad ((\mathbf{S}\cdot\mathbf{N})\,|_{t=0}=-1/c<0),\end{cases}$$

$$\gamma_{(+)}=\begin{cases}\psi_{(\Gamma)} & \text{for } \mathbf{x}\in\partial Q,\quad (\mathbf{S}\cdot\mathbf{N})=(\Omega\cdot\mathbf{n})>0,\\ \psi_{(1)} & \text{for } t=T\quad ((\mathbf{S}\cdot\mathbf{N})\,|_{t=T}=1/c>0),\end{cases}$$

$$\gamma(t, \mathbf{x}, \Omega)=\begin{cases}\gamma_{(-)} & \text{for } \mathbf{r}\in\partial Q_T,\ (\mathbf{S}\cdot\mathbf{N})<0,\\ \gamma_{(+)} & \text{for } \mathbf{r}\in\partial Q_T,\ (\mathbf{S}\cdot\mathbf{N})>0.\end{cases}$$

Taking now into account the notation introduced and denoting by S_0 the operator $S_0=[1/(4\pi)]\int_{\Omega}d\Omega'$, problems (3.35) and (3.36)

*If it is found that $\sigma_{c0}=0$, then via the substitution $\varphi=e^{t\lambda}\tilde{\varphi}$, $\lambda=\text{const}>0$ it is possible to go over from (3.35) to a new equivalent problem of the form (3.35) for the function $\tilde{\varphi}(t, \mathbf{x}, \Omega)$ in which we have $\sigma+\lambda$ instead of σ and with a proper choice of λ the condition $\sigma+\lambda-\sigma_s\geqslant\tilde{\sigma}_{c0}>0$ will be satisfied.

can be written as

$$(\mathbf{S}\cdot\nabla)\ \varphi + \sigma\varphi = \sigma_s S_0\varphi + f, \quad \varphi = \gamma_{(-)} \quad \text{for}$$
$$\mathbf{r} \in \partial Q_T, \ (\mathbf{S}\cdot\mathbf{N}) < 0, \quad (3.37)$$

and

$$-(\mathbf{S}\cdot\nabla)\,\psi + \sigma\psi = \sigma_s S_0\psi + g, \quad \psi = \gamma_{(+)} \quad \text{for}$$
$$\mathbf{r} \in \partial Q_T, \quad (\mathbf{S}\cdot\mathbf{N}) > 0. \quad (3.38)$$

The operators of these problems can be defined on the functions of $H^{(1)}$ that satisfy almost everywhere the boundary values on $\partial Q_{T,-}$ (for problem (3.37)) and on $\partial Q_{T,+}$ (for (3.38)). The existence and uniqueness of solutions of problems (3.37) and (3.38) are sufficiently easy to prove by transforming to a new coordinate system in which one of the axes would be directed along the vector $\mathbf{S}$. Therefore the proof of these results is omitted here and their validity is assumed.

We perform the transformation of problems (3.37) and (3.38). To do this we add and subtract the first equations of (3.37) and (3.38). As a result we arrive at a system

$$(\mathbf{S}\cdot\nabla)\ v + \sigma u = \sigma_s S_0 u + F,$$
$$(\mathbf{S}\cdot\nabla)\ u + \sigma v = \sigma_s S_0 v + G, \quad (3.39)$$

where $u = (\varphi + \psi)/2$, $v = (\varphi - \psi)/2$, $F = (f + g)/2$, $G = (f - g)/2$, the boundary conditions of (3.37) and (3.38) going over into the following:

$$u + v = \gamma_{(-)} \text{ for } \mathbf{r} \in \partial Q_T, \ (\mathbf{S}\cdot\mathbf{N}) < 0,$$
$$u - v = \gamma_{(+)} \text{ for } \mathbf{r} \in \partial Q_T, \ (\mathbf{S}\cdot\mathbf{N}) > 0. \quad (3.40)$$

Determining v from the second equation of (3.39), i.e. assuming

$$v = -B\,((\mathbf{S}\cdot\nabla)\,u - G),$$

where

$$Bw = \frac{1}{\sigma}\,w + \frac{\sigma_s}{\sigma\,(\sigma - \sigma_s)}\,S_0 w, \quad w \in L_2,$$

and substituting into the first equation of (3.39) and the boundary conditions (3.40) we arrive at a problem for determining $u\ (t, \mathbf{x}, \Omega)$:

$$-(\mathbf{S}\cdot\nabla)\ B\ \{(\mathbf{S}\cdot\nabla)\ u - G\} + \sigma u = \sigma_s S_0 u + F,$$
$$u - B\ \{(\mathbf{S}\cdot\nabla)\ u - G\} = \gamma_{(-)} \text{ for } \mathbf{r} \in \partial Q_T, \ (\mathbf{S}\cdot\mathbf{N}) < 0, \quad (3.41)$$
$$u + B\ \{(\mathbf{S}\cdot\nabla)\ u - G\} = \gamma_{(+)} \text{ for } \mathbf{r} \in \partial Q_T, \ (\mathbf{S}\cdot\mathbf{N}) > 0.$$

So if φ and ψ are solutions of problems (3.35) and (3.36), respectively, then the function $u = (\varphi + \psi)/2$ satisfies almost everywhere equa-

tions (3.41). On the other hand, if some function $u\ (t,\ \mathbf{x},\ \Omega)$ satisfies (3.41), then defining

$$\begin{aligned} v &= -B\ \{(\mathbf{S}\cdot\nabla)\ u - G\}, \\ \varphi &= u + v = u - B\ \{(\mathbf{S}\cdot\nabla)\ u - G\}, \\ \psi &= u - v = u + B\ \{(\mathbf{S}\cdot\nabla)\ u - G\} \end{aligned} \tag{3.42}$$

we can see that φ and ψ are solutions of problems (3.35) and (3.36), respectively.

It may be concluded from the foregoing that the original problem (3.35) and its conjugate are equivalent to a single problem, (3.41).

3.3.3. Variational formulation of the problem. To study problem (3.41) we introduce the following Hilbert space:

$$W_1 = \Big\{u:\ (u,\ w)_{W_1} = \int\limits_{Q_T} d\mathbf{r} \int\limits_{\Omega} \Big(\frac{1}{\sigma}\,(\mathbf{S}\cdot\nabla)\,u\cdot(\mathbf{S}\cdot\nabla)\,w + \sigma u w\Big)\,d\Omega$$

$$+ \int\limits_{\Omega} d\Omega \int\limits_{\partial Q_T} |(\mathbf{S}\cdot\mathbf{N})|\,uw\,d\mathbf{r};\ \|u\|_{W_1} = (u,\ u)^{1/2}_{W_1} < \infty\Big\}.$$

We now give a generalized statement of problem (3.41). To do this we perform a scalar multiplication of the basic equation of the problem in L_2 by an arbitrary function $w \in W_1$. If then we carry out an integration by parts according to formula (3.33) taking into account boundary conditions, then we arrive at an equation

$$[u,\ w] = (F,\ w) + (BG,\ (\mathbf{S}\cdot\nabla)\,w) + \int\limits_{\partial Q_T\times\Omega} |(\mathbf{S}\cdot\mathbf{N})|\,\gamma w\,d\mathbf{r}\,d\Omega, \tag{3.43}$$

where

$$[u,\ w] = \int\limits_{\partial Q_T\times\Omega} |(\mathbf{S}\cdot\mathbf{N})|\,uw\,d\mathbf{r}d\Omega + (B\,(\mathbf{S}\cdot\nabla)\,u,\ (\mathbf{S}\cdot\nabla)\,w)$$

$$+ (\sigma u,\ w) - (\sigma_s S_0 u,\ w).$$

Definition. *A generalized solution of problem* (3.41) *is a function* $u \in W_1$ *that satisfies equation* (3.43) *for an arbitrary* $w \in W_1$.

So if u is a solution of (3.41), then that function satisfies (3.43). On the other hand, if (3.43) holds for some function, with $(\mathbf{S}\cdot\nabla) \times B\ \{(\mathbf{S}\cdot\nabla)\ u - G\} \in L_2$ and $B\ \{(\mathbf{S}\cdot\nabla)\,u - G\}|_{\partial Q_T} \in L_2\ (\partial Q_T \times \Omega)$, then carrying out in (3.43) integration by parts and taking into account the fact that the choice of w is arbitrary we see that u satisfies almost everywhere equations (3.41). This definition of a generalized solution is thus an extension of the classical definition.

Going over to it allows us to ignore the questions of the existence of the trace of a function $\Phi = B\{(\mathbf{S}\cdot\nabla)u - G\}$ on ∂Q_T, permits a wider choice of f, g, etc.

Consider some properties of the operator B and of the bilinear form $[u, v]$. It is easy to notice that B is symmetric and positive definite in L_2:

$$(Bv, w) = (v/\sigma, w) + \int_{\partial Q_T} [(\sigma_s \cdot 4\pi)/(\sigma\cdot\sigma_c)(S_0 v)]\,(S_0 w)\,d\mathbf{r} = (Bw, v), \tag{3.44}$$

$$(Bv, v) \geqslant (v/\sigma, v) \geqslant (1/\sigma_1)\,\|v\|^2, \quad v, w \in L_2.$$

Lemma 3.1 [3.17]. *The form $[u, v]$ considered on $W_1 \times W_1$ is symmetric, W_1-bounded and W_1-definite, i.e.*

$$[u, v] = [v, u], \quad |[u, v]| \leqslant c_0 \|u\|_{W_1} \cdot \|v\|_{W_1},$$
$$c_1 \|u\|^2_{W_1} \leqslant [u, u], \tag{3.45}$$

where c_0 and c_1 are positive constants.

Using the properties of $[u, v]$ it is possible to define a Hilbert (energy) space W_A with scalar product $(u, v)_{W_A} \equiv [u, v]$ and norm $[u] = [u, u]^{1/2} < \infty$, which is a completion of W_1 in the metric $[\cdot]$. From (3.45) we find that the spaces W_1 and W_A coincide.

The following theorems are true.

Theorem 3.1 [3.17]. *Let $\gamma \in L_2(\partial Q_T \times \Omega)$, $F \in L_2$, $G \in L_2$. Then for the function $u \in W_1$ to be a generalized solution of problem* (3.41) *it is necessary and sufficient that u minimize in W_A the functional*

$$F(v) = [v, v] - 2(F, v) - 2(BG, (\mathbf{S}\cdot\nabla)v)$$
$$- 2\int_{\partial Q_T^-\times\Omega} |(\mathbf{S}\cdot\mathbf{N})|\gamma v\,d\mathbf{r}\,d\Omega. \tag{3.46}$$

This generalized solution exists, is unique and there is for it an estimate

$$\|u\|_{W_1} \leqslant \text{const}\cdot(\|\gamma\|_{L_2(\partial Q_T^-\times\Omega)} + \|F\| + \|G\|). \tag{3.47}$$

Remark. It is easy to notice that the boundary conditions of (3.41) are natural in the problem under consideration.

We now introduce functions

$$\varphi = u - B\{(\mathbf{S}\cdot\nabla)u - G\}, \quad \psi = u + B\{(\mathbf{S}\cdot\nabla)u - G\} \tag{3.48}$$

which we call generalized solutions of problems (3.35) and (3.36) respectively.

Theorem 3.2 [3.17]. *Let u be a generalized solution of problem*

(3.41). *Then the function* φ *of* (3.48) *satisfies the identity*

$$-(\varphi,\ (\mathbf{S}\cdot\nabla)\,v)+(\sigma\varphi,\ v)=(\sigma_s\ S_0\varphi,\ v)+(f,\ v)$$
$$+\int\limits_{\partial Q_{T,-}\times\Omega}|(\mathbf{S}\cdot\mathbf{N})|\gamma_{(-)}v\,d\mathbf{r}\,d\Omega \qquad (3.49)$$

given an arbitrary $v\in W_1$ *for which* $v|_{\partial Q_{T,+}}=0$.

Theorem 3.3 [3.17]. *If* u *is a generalized solution of* (3.41), *then the function* ψ *of* (3.48) *satisfies the identity*

$$(\psi\,(\mathbf{S}\cdot\nabla)\,v)+(\sigma\psi,\ v)=(\sigma_s S_0\psi,\ v)+(g,\ v)$$
$$+\int\limits_{\partial Q_{T,-}\times\Omega}|(\mathbf{S}\cdot\mathbf{N})|\gamma_{(+)}v\,d\mathbf{r}\,d\Omega \qquad (3.50)$$

given an arbitrary $v\in W_1$ *for which* $v|_{\partial Q_{T,-}}=0$.

Remark. Note that assuming $g=0$, $\psi_{(\Gamma)}=0$ and $\psi_1=0$ results in $\psi\equiv 0$ and solutions of problems (3.35) and (3.41) turn out to be connected by a simple relation $\varphi=2u$.

3.4. APPLICATIONS OF VARIATIONAL PRINCIPLES

As an illustration of possible applications of the variational principle formulated in Sec. 3.3 for the kinetic equation we obtain in this section some approximate mathematical models for describing the unsteady process of particle transport. To this end we use the Ritz method of minimizing functional (3.46) (see Sec. 3.3).

3.4.1. Description of the unsteady particle transport process using a system of ordinary differential equations. We shall minimize functional (3.46) on functions of the form

$$u_M(\Omega,\ t)=u_0(t)+\sum_{i=1}^{3}\Omega_i u_i(t). \qquad (3.51)$$

To find the unknown functions $u_i(t)$ we substitute (3.51) into (3.46) and carry out simple (but rather tiresome) transformations of the terms in (3.46) collecting like terms. As a result we get

$$F(u_0,\ u_1,\ u_2,\ u_3)=\frac{4\pi}{c^2}\int\limits_0^T dt\left\{\left\langle\frac{1}{\sigma_c}\right\rangle\left|\frac{du_0}{dt}\right|^2+\frac{1}{3}\left\langle\frac{1}{\sigma}\right\rangle\sum_{i=1}^{3}\left|\frac{du_i}{dt}\right|^2\right\}$$
$$+4\pi\int\limits_0^T dt\left\{\left(\langle\sigma_c\rangle+\frac{\text{mes }\Gamma}{2}\right)u_0^2+\frac{1}{3}\langle\sigma\rangle\sum_{i=1}^{3}u_i^2\right.$$
$$\left.+2u_0\cdot\sum_{i=1}^{3}u_i(t)\int\limits_\Gamma dx S_0\,(|(\Omega\cdot\mathbf{n})|\cdot\Omega_i)\right.$$

$$+\sum_{i,j=1}^{3} u_i u_j \int_\Gamma dx S_0 \left(|(\Omega\cdot\mathbf{n})|\,\Omega_i\Omega_j\right)\Big\}$$

$$-\frac{8\pi}{c}\Big\{\langle S_0\varphi_0\rangle\, u_0|_{t=0}+\langle S_0\psi_1\rangle\cdot u_0|_{t=T}$$

$$+\sum_{i=1}^{3}\left[\langle S_0\left(\varphi_{(0)}\Omega_i\right)\rangle\cdot u_i|_{t=0}+\langle S_0\left(\psi_{(1)}\Omega_i\right)\rangle\cdot u_i|_{t=T}\right]\Big\}$$

$$-8\pi\int_0^T dt\cdot u_0(t)\Big\{\int_{\Omega\times\partial D_+}\frac{dx\,d\Omega}{4\pi}\,\psi_{(\Gamma)}\,|(\Omega\cdot\mathbf{n})|$$

$$+\int_{\Omega\times\partial D_-}\frac{dx\,d\Omega}{4\pi}\,\varphi_{(\Gamma)}|(\Omega\cdot\mathbf{n})|\Big\}$$

$$-8\pi\int_0^T dt\,\Big\{\langle F_0\rangle\cdot u_0(t)+\sum_{i=1}^{3}\langle F_i\rangle\cdot u_i(t)\Big\}$$

$$-8\pi\cdot\sum_{i=1}^{3}\int_0^T dt\cdot u_i(t)\Big\{\int_{\Omega\times\partial D_+}\frac{dx\,d\Omega}{4\pi}\,|(\Omega\cdot\mathbf{n})|\,\Omega_i$$

$$+\int_{\Omega\times\partial D_-}\frac{dx\,d\Omega}{4\pi}\,\varphi_{(\Gamma)}|(\Omega\cdot\mathbf{n})\,\Omega_i\Big\}$$

$$+\frac{4\pi}{c}(\text{mes } D)\Big\{u_0^2(0)+u_0^2(T)+\frac{1}{3}\sum_{i=1}^{3}\left(u_i^2(0)+u_i^2(T)\right)\Big\}$$

$$+\frac{8\pi}{c}\int_0^T dt\,\Big\{u_0(t)\cdot\frac{d}{dt}\left\langle\frac{G_0}{\sigma_c}\right\rangle+\sum_{i=1}^{3}u_i(t)\,\frac{d}{dt}\left\langle\frac{G_i}{\sigma}\right\rangle\Big\}$$

$$-\frac{8\pi}{c_c}\Big\{u_0(T)\cdot\left\langle\frac{G_0}{\sigma_c}\right\rangle(T)-u_0(0)\left\langle\frac{G_0}{\sigma_c}\right\rangle(0)\Big\}$$

$$-8\pi\sum_{i=1}^{3}\Big\{u_i(T)\left\langle\frac{G_i}{\sigma}\right\rangle(T)-u_i(0)\left\langle\frac{G_i}{\sigma}\right\rangle(0)\Big\},\tag{3.52}$$

where the following notation is introduced:

$$\langle v\rangle = \int_D v\,d\mathbf{x}, \quad S_0 v = [1/(4\pi)] \int_\Omega v\,d\Omega, \quad F_0 = S_0 F, \quad G_0 = S_0 G,$$

$$F_i = S_0(F\Omega_i), \quad G_i = S_0(G\Omega_i), \quad i = 1, 2, 3.$$

Suppose that function (3.51) minimizes functional (3.52) and find the necessary conditions of this. Let $u_{\alpha_j}(t) = u_i(t) + \alpha_i \eta_i(t)$, $i = 0, 1, 2, 3$, where α_i are small numerical parameters and $\eta_i(t)$, are arbitrary, sufficiently smooth functions. Since (under the hypothesis) (3.52) attains its minimum on (3.51), its first variation is zero, which in turn leads to the system

$$\partial F(u_{\alpha_0}, \ldots, u_{\alpha_3})/\partial\alpha_i|_{\alpha_0 = \ldots = \alpha_3 = 0} = 0, \quad i = 0, 1, 2, 3. \tag{3.53}$$

Evaluating the derivatives in (3.53) and taking into account the arbitrariness in the choice of η_i we employ the usual technique of variational calculus to go over from (3.53) to a system of equations in vector matrix form:

$$-\frac{1}{c}\frac{d}{dt}\left[\frac{1}{c}M(t)\frac{d\mathbf{u}}{dt} - \langle \mathbf{G}\rangle\right] + B\mathbf{u}(t) = \langle \mathbf{F}\rangle + \boldsymbol{\gamma} \tag{3.54}$$

with boundary conditions of the form

$$-\frac{1}{c}M(0)\frac{d\mathbf{u}}{dt}(0) + \langle \mathbf{G}\rangle(0) + (\operatorname{mes} D)\,\mathbf{u}(0) = \boldsymbol{\beta}_0, \tag{3.55}$$

$$\frac{1}{c}M(T)\frac{d\mathbf{u}}{dt}(T) - \langle \mathbf{G}\rangle(T) + (\operatorname{mes} D)\,\mathbf{u}(T) = \boldsymbol{\beta}_1, \tag{3.56}$$

where

$$B(t) = \begin{bmatrix} B_{00} & \ldots & B_{03} \\ \cdot & \diagdown & \cdot \\ \cdot & \diagdown & \cdot \\ \cdot & \diagdown & \cdot \\ B_{30} & \ldots & B_{33} \end{bmatrix}, \quad M(t) = \begin{bmatrix} M_{00} & \ldots & M_{03} \\ \cdot & \diagdown & \cdot \\ \cdot & \diagdown & \cdot \\ \cdot & \diagdown & \cdot \\ M_{30} & \ldots & M_{33} \end{bmatrix},$$

$$B_{00} = \langle \sigma_c\rangle + (\operatorname{mes}\Gamma)/2, \; B_{ii} = (1/3)\langle\sigma\rangle + \int_{\partial D} dx\, S_0(\Omega_i^2 |(\Omega\cdot\mathbf{n})|),$$

$$B_{0i} = B_{i0} = \int_{\partial D} dx\, S_0(|(\Omega\cdot\mathbf{n})|\cdot\Omega_i),$$

$$B_{ij} = \int_{\partial D} dx\, S_0(|(\Omega\cdot\mathbf{n})|\Omega_i\Omega_j) \quad \text{for } i \neq j, \; i, j = 1, 2, 3,$$

$$M_{00} = \langle 1/\sigma_c\rangle, \quad M_{ii} = (1/3)\langle 1/\sigma\rangle, \quad i = 1, 2, 3,$$

$$M_{ij} = 0 \quad \text{for } i \neq j, \; i, j = 0, 1, 2, 3,$$

$$\mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{bmatrix}, \quad \langle \mathbf{F} \rangle = \begin{bmatrix} \langle F_0 \rangle \\ \langle F_1 \rangle \\ \langle F_2 \rangle \\ \langle F_3 \rangle \end{bmatrix}, \quad \langle \mathbf{G} \rangle = \begin{bmatrix} \langle G_0/\sigma_c \rangle \\ \langle G_1/\sigma_c \rangle \\ \langle G_2/\sigma_c \rangle \\ \langle G_3/\sigma_c \rangle \end{bmatrix},$$

$$\boldsymbol{\beta}_0 = \begin{bmatrix} \langle S_0 \varphi_{(0)} \rangle \\ \langle S_0 (\varphi_{(0)} \Omega_1) \rangle \\ \langle S_0 (\varphi_{(0)} \Omega_2) \rangle \\ \langle S_0 (\varphi_{(0)} \Omega_3) \rangle \end{bmatrix}, \quad \boldsymbol{\beta}_1 = \begin{bmatrix} \langle S_0 \psi_{(1)} \rangle \\ \langle S_0 (\psi_{(1)} \Omega_1) \rangle \\ \langle S_0 (\psi_{(1)} \Omega_2) \rangle \\ \langle S_0 (\psi_{(1)} \Omega_3) \rangle \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} [1/(4\pi)] \int\limits_{\Omega \times \partial D_+} \psi_{(\Gamma)} |(\Omega \cdot \mathbf{n})| d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_+} \psi_{(\Gamma)} \Omega_1 |(\Omega \cdot \mathbf{n})| d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_+} \psi_{(\Gamma)} \Omega_2 |(\Omega \cdot \mathbf{n})| d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_+} \psi_{(\Gamma)} \Omega_3 |(\Omega \cdot \mathbf{n})| d\mathbf{x}\, d\Omega \end{bmatrix}$$

$$+ \begin{bmatrix} [1/(4\pi)] \int\limits_{\Omega \times \partial D_-} \varphi_{(\Gamma)} |(\Omega \cdot \mathbf{n})|\, d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_-} \varphi_{(\Gamma)} \Omega_1 |(\Omega \cdot \mathbf{n})|\, d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_-} \varphi_{(\Gamma)} \Omega_2 |(\Omega \cdot \mathbf{n})|\, d\mathbf{x}\, d\Omega \\ [1/(4\pi)] \int\limits_{\Omega \times \partial D_-} \varphi_{(\Gamma)} \Omega_3 |(\Omega \cdot \mathbf{n})|\, d\mathbf{x}\, d\Omega \end{bmatrix}.$$

System (3.54), together with the boundary conditions (3.55) and (3.56), is just the desired approximate mathematical model for the description of the unsteady-state transport process. It is obvious that using problems (3.54) to (3.56) it is appropriate to calculate (hoping for an admissible error in this case) approximations to some averaged solutions of problems (3.35) and (3.36) such as

$$(1/\text{mes } D) \langle S_0 \varphi \rangle = (1/\text{mes } D) \int\limits_D d\mathbf{x}\, [1/(4\pi)] \int\limits_\Omega \varphi(\mathbf{x}, \Omega, t)\, d\Omega,$$

$$(1/\text{mes } D) \langle S_0 \psi \rangle = (1/\text{mes } D) \int\limits_D d\mathbf{x}\, [1/(4\pi)] \int\limits_\Omega \psi(\mathbf{x}, \Omega, t)\, d\Omega.$$

To calculate these approximations (denote them by $\varphi_{0,M}$ and $\psi_{0,M}$, respectively) it suffices to replace in (3.48) $u(t, \mathbf{x}, \Omega)$ by $u_M(\Omega, t)$ and perform the operations $\langle\cdot\rangle/\operatorname{mes} D$ and S_0. As a result we get

$$\varphi_{0,M}(t) = u_0(t) - \frac{1}{c}\left(\frac{1}{\operatorname{mes} D}\int\limits_D \frac{d\mathbf{x}}{\sigma_c(\mathbf{x}, t)}\right)\frac{du_0}{dt}(t) + \frac{1}{\operatorname{mes} D}\int\limits_D \frac{d\mathbf{x}}{\sigma_c(\mathbf{x}, t)}\cdot\frac{1}{4\pi}\int\limits_\Omega G(\mathbf{x}, \Omega, t)\, d\Omega, \quad (3.57)$$

$$\psi_{0,M}(t) = u_0(t) + \frac{1}{c}\left(\frac{1}{\operatorname{mes} D}\int\limits_D \frac{d\mathbf{x}}{\sigma_c(\mathbf{x}, t)}\right)\frac{du_0}{dt}(t) - \frac{1}{\operatorname{mes} D}\int\limits_D \frac{d\mathbf{x}}{\sigma_c(\mathbf{x}, t)}\cdot\frac{1}{4\pi}\int\limits_\Omega G(\mathbf{x}, \Omega, t)\, d\Omega. \quad (3.58)$$

To solve numerically problem (3.54) to (3.56) it is also necessary to evaluate the elements of the matrices M and B. Space and surface integrals of the form

$$J_0 = \int\limits_{\partial D\times\Omega} |(\Omega\cdot\mathbf{n})|\, d\mathbf{x}\, d\Omega, \quad J_i = \int\limits_{\partial D\times\Omega} |(\Omega\cdot\mathbf{n})|\,\Omega_i\, d\mathbf{x}\, d\Omega,$$

$$J_{ij} = \int\limits_{\partial D\times\Omega} |(\Omega\cdot\mathbf{n})|\,\Omega_i\Omega_j\, d\mathbf{x}\, d\Omega, \quad i, j = 1, 2, 3,$$

which determine the elements of B and M, can be found using appropriate quadrature formulas. Rather difficult here is the finding of the integrals J_i, J_{ij}, and J_0 whose regions of integration contain a curvilinear surface ∂D and the integrand is the expression for the normal vector $\mathbf{n}$, etc. Notice, however, that in practice the domain D is often a sphere, a cube, a cylinder, etc. It turns out that in these cases J_i, J_{ij}, and J_0 can be easily found in explicit form.

Consider, for example, the following three special cases. If D is a sphere of radius R, then

$$J_0 = 8\pi^2R^2, \quad J_i = 0, \ i = 1, 2, 3, \quad J_{ij} = 0 \text{ for } i \neq j, \quad J_{ii} = 8\pi^2R^2/3, \ i = 1, 2, 3.$$

If $D = \{x, y, z: 0 < x < a, \ 0 < y < b, \ 0 < z < c\}$, then

$$J_0 = 4\pi(bc + ac + ab), \quad J_i = 0, \ i = 1, 2, 3, \quad J_{ij} = 0 \text{ for } i \neq j,$$

$$J_{11} = 2\pi(bc + ac/2 + ab/2), \quad J_{22} = 2\pi(bc/2 + ac + ab/2), \quad J_{33} = 2\pi(bc/2 + ac/2 + ab).$$

If $D = \{x, y, z: \ x^2 + y^2 \leqslant R^2, \ 0 < z < H\}$, then

$$J_0 = 4\pi^2R(H + R), \quad J_i = 0, \ i = 1, 2, 3, \quad J_{ij} = 0 \text{ for } i \neq j,$$

$J_{11} = \pi^2 R\,((3/2)H + R)$, $J_{22} = \pi^2 R\,((3/2)H + R)$,

$J_{33} = \pi^2 R\,(H + 2R)$.

In all these cases the matrix B is diagonal and so problem (3.54) to (3.56) breaks down into four separate one-dimensional problems for an ordinary differential equation of the second order. This fact simplifies the construction of an approximate solution of the entire problem.

Thus we have obtained an approximate mathematical model for the transport equation as a system of second-order differential equations. We used the Ritz method of minimizing the quadratic functional on functions of the form $u(\Omega, t) \cong \varphi_0(t) + \sum_{i=1}^{3} \Omega_i \varphi_i(t)$, i.e. it was assumed that the function which minimize the functional is independent of x. The resulting system of four equations has a symmetric positive definite operator, which makes it possible to use different methods in constructing the difference scheme. In addition, if the region is symmetric (a sphere, parallelepiped or cylinder), then the system breaks down into four independent equations, since it has a diagonal matrix in these cases. This still further simplifies the solution of the problem.

It is seen from the results of numerical experiments in [3.18] that the solution of system (3.54) to (3.56) can be used to approximate averaged solutions of problems. The resulting model has obvious advantages since from an equation in seven variables we have come to a system involving just one variable. Using the variational principle in considering the functional $F(v)$ it is possible to construct also more precise mathematical models for approximation (3.51), (3.52) by increasing the number of known, reasonably chosen basis functions in the expansion $u(\mathbf{x}, \Omega, t) \cong \sum_{i=1}^{M} u_i(t)\,\varphi_i(x, \Omega)$.

3.4.2. The diffusion equation. We shall seek an approximate minimum of $F(v)$ (see (3.46)) on a subspace $\widetilde{W}_A \subset W_A$ consisting of functions $v(t, \mathbf{x})$ that are independent of angular variables. Clearly the function $u(t, \mathbf{x})$ that realizes the minimum exists and is unique. Assuming that $u(t, \mathbf{x})$ has the necessary smoothness and carrying out simple calculations to simplify the form $F(u(t, \mathbf{x}))$ and to find the first functional variation, we arrive at Euler's equation and the boundary conditions satisfied by $u(t, \mathbf{x})$:

$$-\frac{1}{c^2}\frac{\partial}{\partial t}\frac{1}{\sigma_c}\frac{\partial u}{\partial t} - \frac{1}{3}\sum_{i=1}^{3}\frac{\partial}{\partial x_i}\frac{1}{\sigma}\frac{\partial u}{\partial x_i} + \sigma_c u = F_0 - \frac{1}{c}\frac{\partial}{\partial t}\frac{G_0}{\sigma_c} - \sum_{i=1}^{3}\frac{\partial}{\partial x_i}\frac{G_i}{\sigma},$$

$$\frac{1}{3\sigma}\frac{\partial u}{\partial n}+\frac{u}{2}=\frac{1}{\sigma}\sum_{i=1}^{3}n_iG_i+\gamma_{(\Gamma)},\quad \mathbf{x}\in\partial D,\quad t\in(0,\ T),$$

$$\frac{1}{c\cdot\sigma_c}\frac{\partial u}{\partial t}-u=\frac{G_0}{\sigma_c}-S_0\varphi_{(0)},\qquad \mathbf{x}\in D,\quad t=0,$$

$$\frac{1}{c\cdot\sigma_c}\frac{\partial u}{\partial t}+u=\frac{G_0}{\sigma_c}+S_0\psi_{(1)},\qquad \mathbf{x}\in D,\quad t=T,$$

where

$$\gamma_{(\Gamma)}(\mathbf{x},\ t)=\frac{1}{4\pi}\left\{\int\limits_{\Omega\cap(\Omega\cdot\mathbf{n})<0}|(\Omega\cdot\mathbf{n})|\,\varphi_{(\Gamma)}\,d\Omega+\int\limits_{\Omega\cap(\Omega\cdot\mathbf{n})>0}|(\Omega\cdot\mathbf{n})|\,\psi_{(\Gamma)}\,d\Omega\right\},$$

$$F_0=S_0F,\quad G_0=S_0G,\quad G_i=S_0(\Omega_iG),\quad i=1,\ 2,\ 3.$$

Approximations $\tilde{\varphi}_0$ and $\tilde{\psi}_0$ to the functions of the integrated flux $\varphi_0=S_0\varphi$ and to the "integral value" $\psi_0=S_0\psi$ can be determined from the formulas

$$\tilde{\varphi}_0(\mathbf{x},\ t)=u(\mathbf{x},\ t)-\frac{1}{c\cdot\sigma_c(\mathbf{x},\ t)}\frac{\partial u}{\partial t}(\mathbf{x},\ t)+\frac{G_0(\mathbf{x},\ t)}{\sigma_c(\mathbf{x},\ t)},$$

$$\tilde{\psi}_0(\mathbf{x},\ t)=u(\mathbf{x},\ t)+\frac{1}{c\cdot\sigma_c(\mathbf{x},\ t)}\frac{\partial u}{\partial t}(\mathbf{x},\ t)-\frac{G_0(\mathbf{x},\ t)}{\sigma_c(\mathbf{x},\ t)}.$$

If, however, one considers only the original problem for the transport equation with $g=0$, $\psi_{(\Gamma)}=0$ and $\psi_{(1)}=0$, then the function $\Phi(\mathbf{x},\ t)$ approximating $S_0\varphi$ is found from the system of equations

$$-\frac{1}{c^2}\frac{\partial}{\partial t}\frac{1}{\sigma_c}\frac{\partial\Phi}{\partial t}-\frac{1}{3}\sum_{i=1}^{3}\frac{\partial}{\partial x_i}\frac{1}{\sigma}\frac{\partial\Phi}{\partial x_i}+\sigma_c\Phi=f_0-\frac{1}{c}\frac{\partial}{\partial t}\frac{f_0}{\sigma_c}-\sum_{i=1}^{3}\frac{\partial}{\partial x_i}\frac{f_i}{\sigma},$$

$$\frac{1}{3\sigma}\frac{\partial\Phi}{\partial n}+\frac{\Phi}{2}=\frac{1}{\sigma}\sum_{i=1}^{3}n_if_i+2\tilde{\gamma}_{(\Gamma)},\quad \mathbf{x}\in\partial D,\quad t\in(0,\ T),$$

$$\frac{1}{c\cdot\sigma_c}\frac{\partial\Phi}{\partial t}-\Phi=\frac{f_0}{\sigma_c}-2S_0\varphi_{(0)},\qquad \mathbf{x}\in D,\quad t=0,$$

$$\frac{1}{c\cdot\sigma_c}\frac{\partial\Phi}{\partial t}+\Phi=\frac{f_0}{\sigma_c},\qquad \mathbf{x}\in D,\quad t=T,$$

where

$$f_0 = S_0 f, \quad f_i = S_0(\Omega_i f), \quad i = 1, 2, 3,$$

$$\tilde{\gamma}_{(\Gamma)} = [1/(4\pi)] \int\limits_{\Omega \cap (\Omega \cdot \mathbf{n}) < 0} |(\Omega \cdot \mathbf{n})| \, \varphi_{(\Gamma)}(t, \mathbf{x}, \Omega) \, d\Omega.$$

Thus the functions $u(\mathbf{x}, t), \Phi(\mathbf{x}, t) \in \tilde{W}_A$ realizing the minimum of the functionals $F(u)$ and $\tilde{F}(u) = F(u)/4$ on $\tilde{W}_A$ are the solutions to the elliptic equation. The boundary conditions in these problems, having been obtained from the condition of the minima of the functionals, are optimal for this equation.

References

3.1. G. I. Marchuk, *Numerical Methods of Nuclear Reactor Design* (Moscow: Atomizdat, 1958) (in Russian).

3.2. G. I. Marchuk, *Methods of Nuclear Reactor Design* (Moscow: Gosatomizdat, 1961) (in Russian).

3.3. G. I. Marchuk, *Atomnaya Energiya* 11, No. 4: 356-369 (1961) (in Russian).

3.4. G. I. Marchuk, *Methods of Intermediate and Fast Neutron Reactor Design* (Vienna: International Atomic Energy Agency, 1962).

3.5. G. Ya. Rumyantsev, *Atomnaya Energiya* 10, No. 1: 26-34 (1961) (in Russian).

3.6. V. V. Smelov, *Zh. Vychisl. Mat. i Mat. Fiz.* 20, No. 1: 121-132 (1980) (in Russian).

3.7. U. M. Sultangazin, *Methods of Spherical Harmonics and Discrete Ordinates in Problems of Kinetic Transport Theory* (Alma-Ata: Akad. Nauk Kazakh. SSR, 1979) (in Russian).

3.8. G. I. Marchuk and V. I. Lebedev, *Numerical Methods in Neutron-Transport Theory* (Moscow: Atomizdat, 1981) (in Russian).

3.9. G. I. Marchuk and V. V. Orlov, in: *Neutron Physics* (Moscow: Gosatomizdat, 1961): 30 (in Russian).

3.10. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd ed., revised and enlarged (Moscow: Nauka, 1980) (in Russian).

3.11. G. I. Marchuk, *Numerical Solution of Problems in the Dynamics of the Atmosphere and Ocean* (Leningrad: Gidrometeoizdat, 1974) (in Russian).

3.12. V. S. Vladimirov, in: *Trudy Matem. Inst. Akad. Nauk SSSR* 61 (1961) (in Russian).

3.13. V. V. Smelov, *Zh. Vychisl. Mat. i Mat. Fiz.* 19, No. 1: 248-252 (1979) (in Russian).

3.14. S. G. Mikhlin, *Variational Methods in Mathematical Physics*, 2nd ed., revised and enlarged (Moscow: Nauka, 1970) (in Russian).

3.15. A. Sh. Akishev, Cand. Sc. Thesis (Phys.-Math.) (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1983) (in Russian).

3.16. V. I. Gryn', *Zh. Vychisl. Mat. i Mat. Fiz.* 24, No. 1: 75-91 (1984) (in Russian).

3.17. G. I. Marchuk and V. I. Agoshkov, *Symmetrization of the Time-Dependent Transport Equation and Formulation of the Variational Principle* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint No. 222, 1980) (in Russian).

3.18. V. I. Agoshkov and N. I. Brigadirenko, in: *Numerical Methods and Statistical Modelling in Transport Theory* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980): 4-18 (in Russian).

4 Vector Algorithms and Randomization of Monte Carlo Methods

B. A. Kargin, G. A. Mikhailov,
M. A. Nazaraliev, and K. K. Sabelfeld

Basic facts from the theory of vector stochastic algorithms for solving systems of integral equations and algorithms of the method of double randomization for solving problems with random parameters are presented. Theoretical foundations and constructions of the algorithms for solving problems of transport theory taking into account the polarization of light and stochastic heterogeneity of a medium and stochastic problems of elasticity theory are considered.

4.1. VECTOR ALGORITHMS OF THE MONTE CARLO METHOD FOR SOLVING SYSTEMS OF INTEGRAL EQUATIONS OF THE SECOND KIND (VARIANCE, APPROXIMATE OPTIMIZATION)

4.1.1. A number of computational problems can be reduced to the solution of a system of integral equations of the second kind (see, for example, [4.1]-[4.3]). Introducing a suitable discrete-continuous integration measure such a system can be written as a single integral equation, which allows the standard algorithms of the Monte Carlo method to be used [4.4]. It is sometimes more appropriate, however, to use the matrix structure of the system and solve the problem by the Monte Carlo method with vector weights [4.1]-[4.3].

Consider the system of integral equations

$$\varphi_i(x) = \sum_{j=1}^{n} \int_X k_{ij}(x, y)\,\varphi_j(y)\,dy + h_i(x), \quad i = 1, 2, \ldots, m, \qquad (4.1)$$

where $x, y \in X$ and the integral is taken over the Lebesgue measure in a finite-dimensional Euclidean space X. We introduce notation of the form:

$$K(x, y) = \{k_{ij}(x, y)\}_{i,j=1\ 2\ \ldots\ m},$$
$$\Phi(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_m(x))',$$

where the prime designates transposition, i.e. Φ is a column vector. Considering the integral of the matrix as a matrix of integrals we

write the system as

$$\Phi(x) = \int K(x, y)\,\Phi(y)\,dy + \mathbf{H}(x) \tag{4.2}$$

or

$$\Phi = \mathbf{K}\Phi + \mathbf{H}.$$

Equation (4.2) will be considered in L_∞, assuming

$$\|\Phi\| = \operatorname{vrai}\sup_{i,\,x} |\varphi_i(x)|,$$

with

$$\|\mathbf{K}\| \leqslant \operatorname{vrai}\sup_{i,\,x} \sum_j \int |k_{ij}(x, y)|\,dy = q.$$

We define the stationary Markov chain $x_0 = x, x_1, \ldots, x_N$ in X by substochastic transitional density $p(x \to y) = p(x, y)$. Here N is a random length of a chain that terminates with going over from the point x with probability

$$P(x) = 1 - \int p(x, y)\,dy.$$

We shall assume that

$$p(x, y) \neq 0 \quad \text{if} \quad K(x, y)\,\Phi(y) \neq 0.$$

Suppose also that the spectral radius of the operator K_1 obtained from K by replacing $k_{ij} \to |k_{ij}|$, $i, j = 1, 2, \ldots, m$, is less than unity. Then it is possible to show by using the standard techniques of the theory of Monte Carlo methods (see, for example, [4.4] p. 170) that

$$\begin{aligned} &\Phi(x) = M\xi_x, \quad \xi_x = \sum_{n=0}^{N} Q_n \mathbf{H}(x_n), \\ &Q_0 = \{\delta_{ij}\}_{i,\,j=1,\,2,\,\ldots;\,m}, \\ &Q_n = Q_{n-1} K(x_{n-1}, x_n)/p(x_{n-1}, x_n); \end{aligned} \tag{4.3}$$

here

$$K(x_{n-1}, x_n)/p(x_{n-1}, x_n) = \{k_{ij}(x_{n-1}, x_n)/p(x_{n-1}, x_n)\}_{i,j=1,2,\ \ldots,\ m},$$

Q_n may be considered as matrix weights, and M is mathematical expectation.

The Monte Carlo method is used to estimate linear functionals of the form

$$J = (\mathbf{F}, \Phi) = \int \mathbf{F}'(x)\,\Phi(x)\,dx,$$

where $\mathbf{F}'(x) = (f_1(x), f_2(x), \ldots, f_m(x))$, with

$$\|\mathbf{F}\|_{L_1} = \sum_j \int |f_j(x)|\,dx < +\infty.$$

Let the point x_0 be distributed with probability density $\pi(x)$ such that $\pi(x) \neq 0$ if $\mathbf{F}'(x)\,\Phi(x) \neq 0$. Then obviously

$$J = M\left[\frac{\mathbf{F}'(x_0)}{\pi(x_0)}\,\xi_{x_0}\right] = M\sum_{n=0}^{N}\frac{\mathbf{F}'(x_0)}{\pi(x_0)}\,Q_n\mathbf{H}_{\cdot}(x_n)$$

$$= M\sum_{n=0}^{N}\mathbf{H}'(x_n)\,Q_n'\,\frac{\mathbf{F}(x_0)}{\pi(x_0)}.$$

The random vector weights $\mathbf{Q}_n^{(\prime)} = Q_n'\mathbf{F}(x_0)/\pi(x_0)$ are calculated by the formula

$$\mathbf{Q}_n^{(\prime)} = [\mathbf{K}'(x_{n-1}, x_n)/p(x_{n-1}, x_n)]\,Q_{n-1}^{(1)}.$$

Assuming

$$\zeta = [\mathbf{F}'(x_0)\,\xi_{x_0}/\pi(x_0)],$$

we have

$$M\zeta^2 = M[\mathbf{F}'(x_0)\,\xi_{x_0}\xi_{x_0}'\mathbf{F}(x_0)/\pi^2(x_0)].$$

Thus variance $D\zeta$ is determined by the matrix-valued function

$$\Psi(x) = M[\xi_x\xi_x']$$

which is studied below.

4.1.2. In what follows the auxiliary matrix-valued function $\Psi_1(x) = M[\xi_x^{(1)}\xi_x^{(1)\prime}]$, where $\xi_x^{(1)}$ results from replacing $k_{ij} \to |k_{ij}|$ and $h_i \to |h_i|$, is assumed to be finite. This assumption justifies the permutations below in the summation and integration and the finiteness of $\Psi(x)$, just as in the scalar case (see [4.4], p. 153). According to (4.3) we have

$$\Psi(x) = M\sum_{N=0}^{n}Q_n\mathbf{H}(x_n)\cdot\mathbf{H}'(x_n)\,Q_n'$$

$$+ M\sum_{n=0}^{N}Q_n\mathbf{H}(x_n)\Big[\sum_{m=n+1}^{N}Q_m\mathbf{H}(x_m)\Big]'$$

$$+ M\sum_{n=0}^{N}\Big[\sum_{m=n+1}^{N}Q_m\mathbf{H}(x_m)\Big]\mathbf{H}'(x_n)\,Q_n'$$

$$= M_1 + M\sum_{n=0}^{N}Q_n\mathbf{H}(x_n)\,[Q_n(\xi_{x_n} - \mathbf{H}(x_n))]'$$

$$+ M\sum_{n=0}^{N}[Q_n(\xi_{x_n} - \mathbf{H}_{\cdot}(x_n))]\,\mathbf{H}'(x_n)\,Q_n'$$

$$= M_1 + M\sum_{n=0}^{N}Q_n\mathbf{H}(x_n)\,[\Phi'(x_n) - \mathbf{H}^{\ell}(x_n)]\,Q_n'$$

$$+M\sum_{n=0}^{N} Q_n\left[\Phi(x_n)-\mathbf{H}(x_n)\right]\mathbf{H}'(x_n)\,Q_n$$

$$=M\sum_{n=0}^{N} Q_n\left[\mathbf{H}(x_n)\,\Phi'(x_n)+\Phi(x_n)\,\mathbf{H}'(x_n)-\mathbf{H}(x_n)\,\mathbf{H}'(x_n)\right]Q_n'. \tag{4.4}$$

Here standard conditional averaging is used, with a fixed value of the sequence $x_0, x_1, \ldots, x_n$ (see [4.4], p. 153).

The expectation of the nth term in the last sum of (4.4) is expressed as follows:

$$\underbrace{\int\ldots\int}_{n+1}\frac{K(x_0, x_1)}{p(x_0, x_1)}\ldots\frac{K(x_{n-1}, x_n)}{p(x_{n-1}, x_n)}\,G(x_n)\,K'(x_{n-1}, x_n)$$

$$\ldots K'(x_0, x_1)\,dx_0\,dx_1\ldots dx_n, \tag{4.5}$$

where $G=\mathbf{H}\Phi'+\Phi\mathbf{H}'-\mathbf{H}\mathbf{H}'$. It is easy to see that (4.5) is the nth term of the Neumann series for the equation

$$\Psi(x)=\mathbf{H}\Phi'+\Phi\mathbf{H}'-\mathbf{H}\mathbf{H}'$$

$$+\int\left[K(x, y)/p(x, y)\right]\Psi(y)\,K'(x, y)\,dy. \tag{4.6}$$

Thus we have proved the following

Theorem 4.1. *If* $\Psi_1(x)<+\infty$, *then the Neumann series for equation* (4.6) *converges and expresses* $\Psi(x)=M[\xi_x\xi_x']<+\infty$.

4.1.3. We now discuss the conditions for the finiteness of the matrix-valued function $\Psi_1(x)$, which is expressed by the Neumann series for (4.6) after replacing $k_{ij}\to|k_{ij}|$, $h_i\to|h_i|$. The corresponding integral operator will be denoted by $\mathbf{K}_p^{(1)}$. By a direct evaluation we get

$$\|\mathbf{K}_p^{(1)}\|\leqslant\operatorname{vrai}\sup_{i,x}\int_x\left[\Big(\sum_j|k_{ij}(x, y)|\Big)^2\Big/p(x, y)\right]dy$$

$$\leqslant\Big(\operatorname{vrai}\sup_{i,x,y}\sum_j\frac{|k_{ij}(x, y)|}{p(x, y)}\Big)\sup_{i,x}\sum_j\int|k_{ij}(x, y)|\,dy=q_p.$$

Consequently, if $q_p<1$, then $M[\xi_x\xi_x']<+\infty$.

Considering the nth term of the Neumann series for Ψ_1 it is not difficult to extend that condition as follows: $\Psi_1(x)$ is finite if

$$(\operatorname{vrai}\sup_{x,y}[\|K_1(x, y)\|/p(x, y)])\,\rho(K_1)<1,$$

where $\rho(K_1)$ is the spectral radius of the operator K_1 obtained from K after the above substitution. The relations found show that to construct $p(x, y)$ it is expedient to use as uniformly large function $|k_{ij}(x, y)|$ as possible.

An explicit study of the questions of finiteness and of the possibility of minimizing variance for vector algorithms is practically infeasible. A special approximate approach is therefore developed below based on the use of a natural scalar upper limit of a vector estimate. We introduce the notation:

$k(x, y)$ is the norm of the matrix $K(x, y)$,

$$q(x) = \int k(x, y)\, dy,$$

$$h(x) = \Big[\sum_{i=1}^{m} h_i^2(x)\Big]^{1/2} = \| \mathbf{H}(x) \|, \quad f(x) = \| \mathbf{F}(x) \|,$$

$$Q_n = Q_{n-1} k(x_{n-1}, x_n)/p(x_{n-1}, x_n), \quad Q_0 = f(x_0)/\pi(x_0),$$

K is an integral operator with kernel $k(x, y)$,

$$\zeta_0 = \sum_{n=0}^{N} Q_n h(x_n).$$

Theorem 4.2. *Let* $p(x, y) \neq 0$ *for* $k(x, y) \neq 0$ *and* $\pi(x) \neq 0$ *for* $f(x) \neq 0$ *and the spectral radius* $\rho(K) < 1$. *Then*

(i) *there is a unique solution of system* (4.1);

(ii) $M\zeta = (\mathbf{F}, \mathbf{\Phi})$;

(iii) $M\zeta^2 \leqslant M\zeta_0^2$.

Proof. The first and the second statement of the theorem follow from the fact that

$$\rho(\mathbf{K}) \leqslant \rho(\mathbf{K}_1) \leqslant \rho(K) \quad \text{and}$$

$$|\mathbf{F}'(x)\, \mathbf{\Phi}(x)| \leqslant \| \mathbf{F}(x) \| \, \| \mathbf{\Phi}(x) \|.$$

The inequality for the spectral radii follows from the fact that the norm of the integral of the operator does not exceed the integral of the norm. Further, it is obvious that

$$|\zeta| \leqslant \sum_{n=0}^{N} Q_n h(x_n) = \zeta_0.$$

Thus Theorem 4.2 is proved.

So if the kernel $k(x, y)$, for example, is substochastic, then the corresponding algorithm for the direct simulation [4.4] ensures that $M\zeta^2$ is finite. In the general case, if $p(x, y) = k(x, y)/q(x)$, then $M\zeta^2 < +\infty$ provided $\rho(K_q) < 1$, where K_q is an operator with kernel $q(x)\, k(x, y)$. Theorem 4.2 also shows that one may try to reduce $M\zeta^2$ by reducing $M\zeta_0^2$. In particular, if $D\zeta_0 = 0$, then $M\zeta^2 \leqslant (M\zeta_0)^2$.

4.1.4. The above approach can also be used for approximate optimization of weighted estimates that are constructed in solving many

separated integral equations using a single probability model. Note that such estimates are sometimes considered within the framework of the dependent sampling technique [4.4]. It is obvious that this method is especially efficient if calculations are carried out on computer systems with synchronous parallel processors. For the dependent sampling technique the matrix $K(x, y)$ is diagonal and

$$k(x, y) = \max_i | k_{ii}(x, y) | \tag{4.7}$$

If the kernel is independent of the continuous parameter it is natural to consider

$$k(x, y) = \max_\sigma | k(x, y, \sigma) |. \tag{4.8}$$

Numerical experiments for the method of similar trajectories [4.4] have shown that using kernel (4.7) as transitional density of the Markov chain being simulated can, for example, significantly increase the range of σ in which dispersions of weighted estimates are finite.

Note that (4.7) can also be used in the case where $K(x, y)$ is a triangular matrix, which is the case, in particular, for the system of group equations of neutron physics. It is therefore appropriate to use the method in question for constructing the vector algorithm for reactor design which is formulated in [4.2].

4.2. THE METHOD OF DOUBLE RANDOMIZATION. APPROXIMATE MODELS OF RANDOM FIELDS

4.2.1. Let a linear functional equation $L\varphi = f$ be solved by the Monte Carlo method on the basis of simulating a random process whose trajectories will be denoted by ω. This means that random variables $\xi_k(\omega)$ are constructed such that $M\xi_k(\omega) = J_k$, where J_k are the functionals of φ, $k = 1, 2, \ldots, m$, to be computed.

Also suppose that the operator L and the function f are dependent on a random field σ (a random medium in transport theory, random forces in elasticity theory, etc.) and that $\xi_k = \xi_k(\omega, \sigma)$, $J_k = J_k(\sigma)$ and

$$\mathbf{M}[\xi_k(\omega, \sigma \mid \sigma] = J_k(\sigma),$$

with ω and σ being in general dependent.

Consider a problem of computing the quantities

$$J_k = \langle J_k(\sigma) \rangle, \quad R_{kj} = \langle J_k(\sigma) J_j(\sigma) \rangle, \; k, j = 1, \ldots, m,$$

where $\langle\ \rangle$ stands for the mathematical expectation according to the distribution of σ. Such expectations can be evaluated by the following obvious combined method, for example. One numerically

constructs samplings of σ (see, for example [4.4]) and for each sampling an equation $L_\sigma \varphi_\sigma = f_\sigma$ is solved sufficiently exactly by some numerical or analytical method; statistical estimates of the desired quantities are then constructed. This method may be impracticable in solving complicated multidimensional problems, however, because of the exact solution of the equation being very time-consuming. It is sometimes appropriate therefore to employ the double randomization method (see, for example, [4.4]) which here follows from the relations

$$\begin{aligned} \langle J_k(\sigma)\rangle &= \langle M\xi_k(\omega, \sigma)\rangle = M_{(\omega,\sigma)}\xi_k(\omega, \sigma), \\ \langle J_k(\sigma) J_j(\sigma)\rangle &= M_{(\omega_1, \omega_2, \sigma)}[\xi_k(\omega_1, \sigma)\, \xi_j(\omega_2, \sigma)], \end{aligned} \tag{4.9}$$

where ω_1 and ω_2 are conditionally independent trajectories constructed for one sampling of σ and the subscript of the sign of mathematical expectation indicates to what distribution it corresponds. Relations (4.9) show that to estimate J_k it is sufficient to construct just one trajectory for a fixed σ and to estimate R_{kj} it is sufficient to construct two conditionally independent trajectories.

A natural way of optimizing the above randomization method is by the splitting method (see, for example, [4.4]). To estimate J_k it is implemented as follows. For a fixed σ one constructs n conditionally independent trajectories, i.e. a vector $\omega = (\omega_1, \ldots, \omega_n)$, and uses the random variable

$$\zeta_k^{(n)}(\omega, \sigma) = (1/n) \sum_{i=1}^{n} \xi_k(\omega_i, \sigma).$$

instead of $\zeta_k^{(n)}(\omega, \sigma)$. The optimum value of n is calculated from the formula (see, for example, [4.4])

$$n = \sqrt{(a_2/a_1)(t_1/t_2)}, \tag{4.10}$$

where $a_1 = \langle M_\omega^2 \xi_k\rangle - J_k^2$, $a_2 = \langle D_\omega \xi_k\rangle$, t_1 is the mean computer time corresponding to the sampling of σ, and t_2 is the mean time corresponding to the choice of one sampling of ω.

A straightforward evaluation of a_1, a_2, t_1, and t_2 is difficult. It is possible, however, to obtain statistical estimates of variance and time for two values, n_1 and n_2, of the splitting parameter and to solve the corresponding systems of linear equations, i.e. to use the equations

$$\begin{aligned} a_1 &= \frac{1}{n_2 - n_1}(n_2 D\zeta_k^{(n_2)} - n_1 D\zeta_k^{(n_1)}), \\ a_2 &= \frac{n_1 n_2}{n_2 - n_1}(D\zeta_k^{(n_1)} - D\zeta_k^{(n_2)}), \\ t_1 &= \frac{n_2 t^{(n_1)} - n_1 t^{(n_2)}}{n_2 - n_1}, \qquad t_2 = \frac{t^{(n_1)} - t^{(n_2)}}{n_2 - n_1}. \end{aligned} \tag{4.11}$$

It is useful to correlate the sampling of $\zeta_k^{(n_1)}$ and $\zeta_k^{(n_2)}$. Now consider optimization of a randomized estimate of R_{kj} by splitting. It is natural to use in this case the random variable

$$\rho_{kj}^{(n)} = \frac{\sum_{i=1}^{n} \sum_{t=i+1}^{n} [\xi_k(\omega_1, \sigma)\xi_j(\omega_t, \sigma)]}{n(n-1)/2}, \quad M\rho_{kj}^{(n)} = R_{kj}.$$

It turns out that it is possible to use instead of $\rho_{kj}^{(n)}$ another variable, that is numerically equivalent but computationally more convenient. It happens that the following easily verifiable relation for correlation moments is true:

$$K[\zeta_k^{(n)}(\omega, \sigma), \zeta_j^{(n)}(\omega, \sigma)] = K[J_k(\sigma), J_j(\sigma)] + (1/n)\langle K[\xi_k, \xi_j]\rangle.$$

Considering the last relation also for $n = 1$ and solving the resulting system of equations we get

$$K_{jk} = K[J_k(\sigma), J_j(\sigma)] = (nK[\zeta_k^{(n)}, \zeta_j^{(n)}] - K[\zeta_k^{(1)}, \zeta_j^{(1)}])/(n-1). \tag{4.12}$$

It is easy to see that substituting on the right-hand side of (4.12) statistical estimates for the covariances gives an estimate $\tilde{K}_{kj}^{(n)}$ of K_{jk}, which coincides numerically with the estimate constructed using $\rho_{kj}^{(n)}$. Note that when using (4.12) the mean time of calculations, as in the standard variant of splitting, is expressed by a formula of the form $T^{(n)} = T_1 + nT_2$.

An exact expression of variance $D\rho_{kj}^{(n)}$ is very cumbersome. It is clear, however, that for sufficiently large n there is an approximate equation

$$D\tilde{K}_{kj}^{(n)} \approx A_1 + A_2/n.$$

Hence, in evaluating K_{jk} it is also appropriate to optimize splitting by formulas of the form (4.10), (4.11). Note that the question has been left open concerning the choice of the value of n optimal in some sense for the evaluation of the totality of K_{jk}, $j, k = 1, \ldots, m$.

For randomized algorithms of the Monte Carlo method to be implemented it is necessary to use sufficiently economical ways of simulating random fields. Further we consider two classes of approximate models of random fields whose cost of implementation grows with the degree of approximation to the limiting fields with absolutely continuous finite-dimensional distributions (limiting fields are continuous in the Gaussian case).

4.2.2. We first consider the well-known method of constructing a special random process $\xi(t, k)$ with an arbitrary function of one-dimensional distribution $F_\xi(x)$ and a convex normalized correla-

tion $k(t)$ such that

$$k'(t) \leqslant 0, \quad k''(t) \geqslant 0, \quad |k'(0)| < +\infty, \quad k(+\infty) = 0. \tag{4.13}$$

Under conditions (4.13) the functions

$$f(t) = -k''(t)/k'(0), \quad f_1(t) = -k'(t), \quad t \geqslant 0$$

are distribution densities and determine Palm's stationary point flux

$$\tau_k = \sum_{i=1}^{k} \eta_i, \quad k = 1, 2, \ldots, \quad \tau_0 = 0,$$

where $\{\eta_i\}$ are independent nonnegative random variables, with η_1 distributed with density $f_1(t)$ and η_i, $i \geqslant 2$, with density $f(t)$. For such a flux the probability of the zero of points in an interval of length t is $P(0, t) = k(t)$.

The process $\xi(t, k)$ in the interval $(0, T)$ is constructed as follows:

(i) a sequence $\{\tau_k\}$ is simulated until the first exit from $(0, T)$,

(ii) for every interval (τ_{i-1}, τ_i), $i = 1, 2, \ldots$, it is assumed that $\xi(t) \equiv \xi_i$, where $\{\xi_i\}$ are independent random variables with distribution function $F_\xi(x)$.

On the basis of Palm's flux it is also possible to construct a simple model of an isotropic random field as follows [4.5]:

(i) Palm's flux $\{\tau_k\}$ is constructed with $P(0, t) = k(t)$ along the isotropic direction ω in the range of D;

(ii) for each of the slabs bounded by the planes perpendicular to ω at the points of $\{\tau_k\}$ an independent value of ξ is chosen according to $F_\xi(x)$.

It is natural to call this model an "isotropic layered" model. For it

$$k(r) = c_n \int_0^{\pi/2} k(r \cos \vartheta) \sin^{n-1} \vartheta \, d\vartheta, \quad r = |\mathbf{r}|.$$

For example, if k is absolutely monotonic and $n = 3$, then

$$k(r) = \int_0^1 \int_\Lambda \exp(-\lambda r v) \, \mu(d\lambda) \, dv = \int_\Lambda [(1 - \exp(-\lambda r))/(\lambda r)] \mu(d\lambda).$$

Distribution of the isotropic layered field is unnatural for a number of applications, but in some cases it can be improved using a simple modification, which is described below.

Suppose that the distribution of ξ is divisible without limit, i.e.

for every natural m we have the representation

$$\xi = \xi_i^{(m)} + \ldots + \xi_m^{(m)},$$

where $\xi_i^{(m)}$ are independent and equally distributed. Let

$$\zeta_m(t) = \sum_{i=1}^{m} \xi_i^{(m)}(t, k), \tag{4.14}$$

where $\xi_i^{(m)}(t, k)$ are independent samplings of the process $\xi^{(m)}(t, k)$ with the same one-dimensional distribution as $\xi_i^{(m)}$ has. It is clear that the function of the one-dimensional distribution of the process $\xi_m(t)$ is $F_\xi(t)$ and that the correlation function is $k(t)$. Samplings of $\zeta_m(t)$ are improved compared to $\xi(t, k)$ in that they are close to continuous functions, $\zeta_m(t)$ taking constant values in smaller regions and those values being dependent.

Now consider the convergence of processes $\zeta_m(t)$. It has been established in [4.5] that the corresponding finite-dimensional distributions are weakly convergent to the distributions whose totality satisfies the consistency conditions (this is also true for the fields), the limiting distributions being absolutely continuous if so is $F_\xi(x)$. Using the Chentsov-Kolmogorov criterion [4.6], [4.5] proves

Theorem 4.1. *If* $|k''(t)| < c < +\infty$ *in* $[0, T]$ *and* $\xi \geqslant 0$, *then processes* $\zeta_m(t)$ *are weakly (with respect to the metric D) convergent to the process given by the limiting finite-dimensional distributions, with the correlation function* $k(t)$ *and function* $F_\xi(x)$ *of a one-dimensional distribution.*

It is sufficiently clear that this theorem can be extended to fields by means of an appropriate generalization of the Chentsov-Kolmogorov criterion. In any case the finite-dimensional distributions of the model fields we have considered are improved by summation of the type (4.14). Using the layered isotropic model one can thus construct fields with sufficiently natural samplings. Note that in using random medium models of the type (4.14) in radiation transport theory it suffices to choose a value of m such that the characteristic dimension of the constancy range of $\zeta_m(\mathbf{r})$ is less than the free length of a quantum of radiation.

Note another useful property of the above random field models: they allow fields with different one-dimensional distributions to be simulated dependently (on a single "Palm foundation" corresponding to a given function $k(t)$). It was shown by sampling such models that non-Gaussian fields may radically differ from Gaussian ones in the presence of regions of practically constant value with diameters substantially exceeding the scale of correlation. We may regard this fact as exhibiting the property of the maximality of the entropy of a Gaussian distribution with given covariances.

4.2.3. Next we consider a randomized spectral model of a Gaus-

sian field. The problem is posed of an approximate simulation of a random Gaussian field with a correlation function $k(x)$ for which the following spectral decomposition is given

$$k(x) = \int \cos(\lambda x)\, p(\lambda)\, d\lambda, \tag{4.15}$$

where $p(\lambda)$ is the spectral density, $\lambda \in \Lambda = R_n$. Note that representation (4.15) exists, in particular, if $\int | k(x) | dx < +\infty$. For simplicity it will be assumed that

$$M_\xi(x) = 0, \quad D_\xi(x) = 1.$$

We partition the space Λ into $\Lambda_1, \ldots, \Lambda_m$ and assume that the random points $\lambda_1, \ldots, \lambda_m$ are distributed in the parts according to the densities

$$p_k(\lambda) = p(\lambda) \Big/ \int_{\Lambda_k} p(\lambda)\, d\lambda, \quad \lambda \in \Lambda_k.$$

Then by virtue of (4.15) we have

$$k(x) = M \sum_{k=1}^{m} p_k \cos(\lambda_k x), \quad p_k = \int_{\Lambda_k} p(x)\, dx.$$

From this we obtain the following procedure for constructing a random field with a given correlation function:

(i) choose the values $\lambda_1, \ldots, \lambda_m$,

(ii) construct a sampling of the field from the formulas

$$\xi_m(x) = \sum_{k=1}^{m} p_k^{1/2} [\xi_k \sin(\lambda_k x) + \eta_k \cos(\lambda_k x)], \tag{4.16}$$

where $\{\xi_k, \eta_k\}$ are jointly independent standard Gaussian variables.

The conditional one-dimensional distribution of field (4.15) for fixed values $\lambda_1, \ldots, \lambda_m$ is standard Gaussian. Hence the absolute one-dimensional distribution of (4.15) is also standard. Similar considerations show that multi-dimensional distributions of field (4.15) are not normal and that it is not ergodic, but these disadvantages may grow weaker as $m \to \infty$ and with sufficiently uniform "refinement" of the partition of the space Λ.

Now we consider the convergence of a field $\xi_m(x)$ to a Gaussian field with a given spectral density $p(\lambda)$. As pointed out above, this convergence is related to the character of the partition of Λ into Λ_k. Denote by V_k the volume of a Λ_k and by d_k the diameter of Λ_k. We consider below partitions of $\Lambda \setminus \Lambda_m$ uniform in volume and such that $d_k \leqslant c V_k^{1/n}$, $k = 1, 2, \ldots, m-1$, with $V_k \to 0$ as

$m \to \infty$. This partition can obviously be obtained by using, for example, a uniform rectangular grid with a spacing of order $m^{-1/n}$ in each axis. The following theorem is proved in [4.5].

Theorem 4.2. *Let* $\xi_m(x)$ *be fields given by formula* (4.16) *for* $|x| < R$, $\Lambda_m = \{\lambda: |\lambda| > l_m\}$ *and let* $d_k < c l_m \cdot m^{-1/n}$, $k = 1, \ldots, m-1$. *Then*

(i) *if* $l_m m^{-1/n} \to 0$ *and* $l_m \to \infty$ *as* $m \to \infty$ *we have*

$$M \int\limits_{|x|<R} [\xi_m(x) - \xi(x)]^2 \, dx \underset{m \to 0}{\to} 0,$$

where $\xi(x)$ *is a Gaussian field with a spectral density* $p(\lambda)$;

(ii) *if for some* $\varepsilon > 0$,

$$\int |\lambda|^{\varepsilon} p(\lambda) \, d\lambda < +\infty,$$

for $l_m = c \cdot m^{2/[n(2+\varepsilon)]}$ *there is an estimate*

$$M \int\limits_{|x|<R} [\xi_m(x) - \xi(x)]^2 \, dx \leqslant c_m^{-2\varepsilon/[n(2+\varepsilon)]}.$$

By Chebyshev's inequality Theorem 4.2 gives the convergence in probability to zero of the L_2-deviation of the field $\xi_n(x)$ from the Gaussian field $\xi(x)$ with a given spectral density. It is easy to obtain a similar convergence of the derivatives of $\xi_n(x)$ provided there are appropriate moments of spectral density. Further, using the imbedding theorems one can formulate [4.5] statements about the convergence in probability to zero of a uniform deviation of $\xi_m(x)$ from $\xi(x)$.

To conclude, Theorem 4.2 and its consequences may be useful in decreasing the cost of the double randomization method when solving particular stochastic problems with a desired accuracy.

4.3. SOLVING STOCHASTIC PROBLEMS OF ELASTICITY THEORY

4.3.1. As pointed out in Sec. 4.2, to construct an efficient algorithm for solving stochastic problems, which is based on the principle of double randomization, requires constructing a statistical estimate for the corresponding determinate problem. Before proceeding to the construction of such estimates we briefly discuss some features of the Monte Carlo method for solving boundary-value problems in determinate formulation. Here it is convenient to start with classifying the solution methods and classes of the boundary-value problems under consideration. This will allow us later on to formulate precisely a method for the corresponding class of problems

and to justify clearly the advantages of the proposed methods for the defined range of problems.

We start with the solution methods. Suppose it is necessary to find at a point $x_0 \in G \subset R^3$ a solution of a well-posed boundary-value problem for an equation (or system) of elliptic type

$$Lu(x) = f(x), \quad x \in G, \tag{4.17}$$

$$B_j u|_{\partial G} = \psi_j(t), \quad j = 1, \ldots, m-1, \quad t \in \partial G, \tag{4.18}$$

where L is the corresponding differential operator and $\{B_j\}$ is the family of operators whose values are prescribed by (4.18). The general scheme of the Monte Carlo method for this problem is as follows. Construct a random process ξ_x and the functional F so that

$$u(x) = M_{\xi_x} F[\xi_x]. \tag{4.19}$$

If the equality in (4.19) is exact, the random estimate $F[\xi_x]$ is said to be unbiased; if the equality is approximate, i.e. if

$$u(x) = M_{\xi_x} F[\xi_x] + \varepsilon, \tag{4.20}$$

then the estimate $F[\xi_x]$ is said to be ε-biased. For conditions under which the probability representation (4.19) is possible see, for example, [4.7]. It is clear that one can construct a set of representations (4.19), (4.20). For example, if ξ is the Markov chain:

$$\xi_{k+1} = \xi_k + d_k \omega_k, \quad \xi_0 = x, \tag{4.21}$$

where ω_k is a random isotropic vector which is uniformly distributed over a unit sphere with center at a point ξ_k, $S(\xi_k, d_k) \subset G$, then we have a class of methods of "random walk over spheres". For example, for the Dirichlet problem

$$\Delta u = 0, \qquad x \in G, \tag{4.22}$$

$$u|_\Gamma = \psi(t), \quad t \in \Gamma = \partial G, \tag{4.23}$$

representation (4.20) is in this class of the form

$$u(x) = M[\psi(x^*)] + O(\omega(\varepsilon)), \tag{4.24}$$

where x^* is the boundary point closest to the terminating point ξ_N of the process of "random walk over spheres" in the band $\partial G_\varepsilon = \{x \in G: |x - t| < \varepsilon, t \in \partial G\}$, and $\omega(\varepsilon)$ is the modulus of continuity of $u(x)$ in ∂G_ε. If, however, the states of the Markov chain $\{\xi_n\}$ are points of the boundary ∂G, then there is a "random walk over the boundary" [4.7] with (4.24) assuming the form

$$u(x) = M\left\{\sum_{i=0}^{N} A_i \psi(\xi_i)\right\} + O(\delta), \tag{4.25}$$

where

$$\delta = \operatorname*{osc}_{t \in \partial G} \psi(t) \cdot q^N, \quad q = \text{const} < 1,$$

A_i are known constants [4.7], and the random points $\xi_i \in \partial G$ are defined as follows: ξ_0 is the point of the intersection of a ray of isotropic direction from a point x and the boundary ∂G

$$\xi_k = \xi_{k-1} + r_k \omega_k, \quad k = 1, \ldots, N,$$

where

$$r_k = | \xi_k - \xi_{k-1} |, \quad \xi_k \in \partial G.$$

There are also other random processes solving problem (4.22), (4.23) approximately or exactly, such as the process of a random walk over a grid and the Wiener process, but here it is reasonable to restrict ourselves to the two processes described above.

We now proceed to the class of problems (4.17), (4.18) being solved. Scalar random estimates of the form (4.19), (4.20), (4.24), (4.25) are constructed for second-order equations with the Dirichlet, Neumann conditions or conditions of the third kind, with the method of "random walk over spheres" giving a solution of the interior Dirichlet problem at one (or several) points. The method of random walk over a boundary, which uses the boundary integral equations of potential theory, allows one to solve both interior and exterior problems with the Dirichlet, Neumann conditions and conditions of the third kind. In this case, however, there is no longer such a natural probability interpretation as in the case of "random walk over spheres" and one has to specially construct a uniformly and absolutely convergent series of iterations of a boundary integral operator. But if one solves an equation of a higher order or a system of equations of elliptic type, then vector random estimates are constructed. (For a general approach to constructing such vector estimates see [4.8].) That approach will be considered in detail in the next subsection using an example of the biharmonic problem. Vector estimates on the trajectories of the "process of random walk over spheres" are constructed in exactly the same way as in the scalar case if the spectral radius of the corresponding matrix-integral operator is less than unity. The study of variance becomes more complicated in that case [4.3]. But if one constructs vector estimates determined on the process of random walk over a boundary an extra difficulty arises, that of calculating singular integrals of potential theory which exist in the sense of principal value. These questions are beyond the scope of the present paper, however.

In conclusion we list the basic features of the Monte Carlo method that account for its increasingly intensive use in solving boundary-

value problems in both stochastic and determinate formulations.

1. The algorithms are easy to implement, which allows one to solve dynamically multidimensional problems with complex geometry.

2. The solution and its derivatives are computed simultaneously and with the same order of accuracy. This is especially important in solving higher-order equations where it is often required to calculate higher derivatives of the solution (for example, the intersecting forces in the bending problem).

3. Solving problems of even rather high dimensions does not require large computer memory.

4. The algorithms allow one to find a solution and its derivatives at the desired point or in a given set of points, which is important, for example, in solving a number of extremal problems (finding a maximum deflection, maximum stresses, etc.).

5. One can solve efficiently exterior multidimensional boundary-value problems with practically arbitrary boundary conditions which present certain difficulty to other numerical methods.

6. The method of statistical simulation is practically the only numerical method suitable for solving boundary-value problems with actual stochastic parameters, such as a random boundary and random coefficients, where it is required to calculate certain statistical characteristics of a solution and its derivatives.

We now proceed to construct the simplest vector random estimate for the solution of the biharmonic equation.

4.3.2. Consider the bending problem for a plate $G \cup \partial G$ whose normal deflection satisfies the equation

$$\Delta\Delta u\,(x) = q\,(x,\ y)/D, \quad x \in G \subset R^2, \tag{4.26}$$

and the boundary conditions

$$u|_{\partial G} = \varphi_1, \quad \Delta u|_{\partial G} = \varphi_2. \tag{4.27}$$

Here $q\,(x, y)$ is the intensity of normal loading, $D = Eh^3/[12\,(1 - \sigma^2)]$, E is Young's modulus, σ is the Poisson constant of the material of the plate and $2h$ is the plate thickness.

Let $x_0 \in G$ be the point where a solution is sought. Then for the solution of problem (4.26), (4.27) there is a probability representation

$$u\,(x_0) = M_{x_0}\left\{\varphi_1\,(x_\tau) - \tau\varphi_2\,(x_\tau) + \int_0^\tau [tq\,(x_t)/D]\,dt\right\}, \tag{4.28}$$

where τ is the moment of the first output from G of the Wiener process $w\,(t)$: $w\,(0) = x_0$, $w\,(\tau) = x_\tau \in \partial G$, $w\,(t) = x_t \in G$ and M_{x_0} is the symbol of mathematical expectation in the space of all such $w\,(t)$.

It is in principle possible to use the Monte Carlo method for a straightforward calculation of the Wiener integral (4.28), but this would require a time-consuming procedure of simulating τ [4.3]. We therefore develop here a different approach to the solution of this problem based on constructing a vector random estimate. Consider first the homogeneous equation (4.26). Let $S(x_0, r)$ be a circle with centre at a point x_0 of an arbitrary radius r such that $S(x, r) \subset G$. (It is usual to consider [4.9] the case of a maximum r denoted by R. Thus $S(x, R)$ is a circle of maximum radius with centre at a point x contained entirely in $G \cup \partial G$.)

Let $\{\xi_0, \xi_1, \ldots, \xi_N\}$ be an "ε-spherical process", i.e. a sequence of random points such that $\xi_0 = x$, $\xi_1 = \xi_0 + \omega_1 r_1, \ldots, \xi_N = \xi_{N-1} + \omega_N r_N$, where $\{\omega_1, \ldots, \omega_N\}$ is a sequence of independent vectors uniformly distributed over the circles $S(\xi_0, 1)$, $S(\xi_1, 1), \ldots, S(\xi_{N-1}, 1)$ respectively, with $\xi_N \in \partial G_\varepsilon$ being the first terminating point. The boundary point closest to the terminating is denoted by t_N. The "process of random walk over maximum spheres" well represented in the literature [4.9] is a sequence of random centres of circles $S(\xi_0, R_1)$, $S(\xi_1, R_2), \ldots, S(\xi_{N-1}, R_N)$, where $\xi_k \in S(\xi_{k-1}, R_k)$, $k = 1, \ldots, N$. It is clear that $N \to \infty$ as $\varepsilon \to 0$, but, with ε fixed, the average number of steps till a discontinuity occurs is of the order $O(|\ln \varepsilon|)$ [4.9]. If $r_k < R_k$, with $r_k \sim \varepsilon$, then the average number of steps is known [4.9] to be $O(r^{-2})$ regardless of the dimension of the space.

We call

$$C^{(k)}(r_k) = \begin{pmatrix} 1 & -r_k^2/4 \\ 0 & 1 \end{pmatrix} \tag{4.29}$$

the transition matrix of a vector random estimate for problem (4.26), (4.27) and introduce the following notation: $u(x) = v_1$, $\Delta u(x) = v_2$, $\varphi = (\varphi_1, \varphi_2)^T$; then the vector random estimate $\xi = (\xi_1, \xi_2)^T$ for $(v_1, v_2)^T$ is determined as follows:

$$\xi(x) = \prod_{k=1}^{N} C^{(k)}(r_k)\, \varphi(t_N). \tag{4.30}$$

The equation

$$\mathbf{v}(x) = M\xi(x) + O(\omega(\varepsilon)) \tag{4.31}$$

follows from the generalized mean value theorem for a biharmonic function (see [4.9] and Theorem 4.2 of the preceding section. Estimate (4.30) has finite (at $\varepsilon > 0$) variance and the algorithm based on (4.31) has the cost $O(|\ln \varepsilon|/\varepsilon^2)$ (for the case of maximum r_k).

Before proceeding to the main task, that of solving equation (4.26) with a random field of loads $q(x, y)$, we construct a random estimate which takes into account the right-hand side of equation

(4.26). Here we discuss two ways for such allowance. The first is based on the fact that under zero boundary conditions

$$u(x_0) = [1/(8\pi)] \int\limits_{K(x_0,\, r)} \{r^2 - z^2 [\ln(r/z) + 1]\} f(r')\, dr',$$
$$\Delta u(x_0) = [1/(2\pi)] \int\limits_{K(x_0,\, r)} \ln(r/z) f(r')\, dr', \tag{4.32}$$

where $z = |x_0 - r'|$, $K(x_0, r)$ is a circle of radius r with centre at a point x_0 and $f = q/D$. Consequently, if we estimate (4.32) one random node in every circle $K(\xi_i, r_{i+1})$, then the vector estimate ξ_1 for the solution of problem (4.26), (4.27) will assume the form

$$\xi_1 = \xi + \eta, \tag{4.33}$$

where

$$\eta = \sum_{i=1}^{N} \eta_i,$$

$\eta_i = (\eta_i^{(1)}, \eta_i^{(2)})^T$ is a vector random estimate of integrals in (4.32) in a circle $K(\xi_{i-1}, r_i)$. The simplest estimate for η_i is of the form

$$\eta_i^{(1)} = \frac{r_i^2 - z^2 [\ln(r_i/z) + 1]}{4} \frac{r_i^2}{4 \ln(r_i/z)} f(x_0 + \rho\omega),$$

$$\eta_i^{(2)} = -\frac{r_i^2}{4} f(x_0 + \rho\omega),$$

where ρ is a random variable distributed over the interval $[0, r_i]$ with density $4z \ln(r_i/z)\, r_i^2$ and ω is the unit isotropic vector.

Another method is more convenient for solving problem (4.26), (4.27) with a random right-hand side and zero boundary values.

It may be assumed without loss of generality that

$$\int\limits_G f(x)\, dx = 1.$$

Then

$$u(x_0) = \int\limits_G u_\delta(x_0, y) f(y)\, dy, \tag{4.34}$$

where u_δ is a solution of the problem

$$\Delta\Delta u_\delta(x_0, y) = \delta(x_0 - y),$$
$$u_\delta|_{\partial G} = \Delta u|_{\partial G} = 0. \tag{4.35}$$

The solution $u_\delta(x_0, y)$ will be sought in the orm

$$u_\delta(x_0, y) = V(x_0, y) + w(x_0, y), \tag{4.36}$$

where $V(x_0, y)$ is a fundamental solution of the biharmonic equation:

$$V(x_0, y) = [1/(8\pi)] \mid x_0 - y \mid^2 \ln \mid x_0 - y \mid. \tag{4.37}$$

Then we find w from the condition

$$\Delta\Delta w(x_0, y) = 0, \quad w|_{x_0 \in \Gamma} = -V|_{x_0 \in \Gamma}, \quad \Delta w|_{x_0 \in \Gamma} = -\Delta V|_{x_0 \in \Gamma}. \tag{4.38}$$

From (4.34) we have

$$u(x_0) = Mu_\delta(x_0, y), \tag{4.39}$$

where averaging (mathematical expectation M) is made in the space of trajectories $\{\xi_k\}$ starting from the point x_0 and contributing $u(x_0, y)$ to the solution from the δ-source which is located at a point y (y is a random point distributed in a region G with density f).

So in view of the fact that the function $u_\delta(x_0, y)$ is symmetrical the second algorithm may be formulated as follows. A random point y is chosen in G with density f. From that point we let out a trajectory $\{\xi_1, \ldots, \xi_N\}$ and calculate on it the contribution of the δ-source located at the point x_0. Then a point y is chosen in G with density f and the contribution of the δ-source located at a point x_0 is calculated on the second trajectory, etc., and the result is averaged over a sufficiently rich ensemble of such trajectories. Note that the algorithm gives a solution simultaneously in any set of points where δ-sources are located.

We now discuss calculating derivatives for problem (4.26), (4.27) under homogeneous boundary conditions. The problem is set as follows. On the same trajectory $\{\xi_0, \ldots, \xi_N\}$ where a solution $u(x)$ is constructed it is necessary to construct random estimates of differential operators of the solution at given points. It is easy to see from representation (4.39) that the last algorithm will also work in estimating any differential operator of the solution Lu. Indeed, it is seen from (4.39) that it suffices to construct an estimate for $L_x u_\delta(x, y)$. But

$$L_x u_\delta(x, y) = L_x V(x, y) + L_x w(x, y), \tag{4.40}$$

so it remains to construct an estimate for $L_x w(x, y)$. To do this it is sufficient to consider a random estimate for $w(x, y)$, η_w

$$\eta_w(x, y) = -V(x, y)|_{y \in \Gamma} + [(1/4) \sum_{j=1}^{N} r_j^2] \Delta V(x, y)|_{y \in \Gamma} \tag{4.41}$$

and apply the differential operator L_x to it. This follows from the boundedness of the region and the form of $V(x, y)$. Equation (4.41) uses the symmetry of the function w with respect to x and y. So the

algorithm for calculating the differential operator $Lu(x)$ of the solution of problem (4.26), (4.27) is as follows. Locate a δ-source at a point x. (If it is necessary to calculate simultaneously the derivatives at many points $x_1, \ldots, x_m$, then also locate a δ-source at each of these points.) Choose in G a random point y from the density f and construct on the chain $\xi_0 = y, \xi_1, \ldots, \xi_N$ the estimate

$$\eta(x) = L_x V(y, x) - L_x V(y, x)|_{y\in\Gamma} + \left[(1/4)\sum_{j=1}^{N} r_j^2\right] \Delta L_x V(y, x)|_{y\in\Gamma}. \quad (4.42)$$

It is clear that by virtue of

$$Lu(x_i) = M\eta(x_i), \quad i = 1, \ldots, m,$$

it is possible to estimate $Lu(x)$ simultaneously at all points $x_1, x_2, \ldots, x_m$ on a single trajectory.

4.3.3. Now we discuss calculating statistical characteristics of a solution and differential operators when the right-hand side f in (4.26) is a homogeneous random field.

Let

$$K_f(r) = \langle f(x + r) f(x) \rangle$$

be a correlation function assumed to be equal to (see [4.10])

$$K_f(r) = \sigma_0^2 e^{-\alpha|r|}, \quad (4.43)$$

where σ_0^2 is the dispersion, $r_0 = \alpha^{-1}$ is the correlation scale, and the spectral density corresponding to (4.43) is

$$S(\omega) = (\sigma_0^2/\pi)[\alpha(\alpha^2 + \omega^2)]. \quad (4.44)$$

It is assumed for simplicity that $\varphi_1 = \varphi_2 = 0$. We shall first describe the algorithm for computing the correlation function $K_u(x_1 - x_2)$ of the solution of problem (4.26), (4.27). We shall use the second variant of taking into account the right-hand side based on relations (4.34) to (4.39). We choose a random point y in G with some density p. A trajectory $(\xi_1, \ldots, \xi_N)$ is let out of that point and the contributions of δ-sources located at points x_1 and x_2 are calculated on it: $\eta = [f^2(y)/p^2]\, u_\delta(x_1, y)\, u_\delta(x_2, y)$ for a fixed implementation of f. This process is implemented repeatedly and the result is averaged: $K_u(x_1 - x_2) = M\eta$. It is essential to note that such random estimates can be constructed on a single trajectory to assess arbitrary statistical moments of any differential operators of a solution, which follows from (4.40) and (4.41).

Here we give an estimate for calculating statistical moments of internal force factors. Let

$$L_x u = M_x u = -D(\partial^2/\partial x^2 + \nu\partial^2/\partial y^2)\, u(x, y). \quad (4.45)$$

Then for a fixed implementation of f according to (4.41) it is only necessary to calculate

$$M_x V, \quad -M_x V(x, y)|_{y\in\Gamma}, \quad -M_x \Delta V(x, y)|_{y\in\Gamma}.$$

Simple calculations yield

$$M_x V = -D\left\{\frac{(x_1-x_2)^2}{4\pi r^2} + \frac{2\ln r+1}{8\pi} + \nu\left[\frac{(y_2-y_1)^2}{4\pi r^2} + \frac{2\ln r+1}{8\pi}\right]\right\}, \tag{4.46}$$

$$M_x \Delta V = -D\left[-\frac{x_2-x_1}{\pi r^4} + \frac{1}{2\pi r^2} - \nu\frac{(y_2-y_1)^2}{\pi r^4} + \frac{\nu}{2\pi r^2}\right]. \tag{4.47}$$

Here the point (x_1, y_1) is chosen in a region with density p and (x_2, y_2) is the point where (4.45) is estimated. It remains to substitute (4.46) and (4.47) into (4.32).

In conclusion we list the results of the numerical experiments for $\sigma_0 = 1$, $\alpha = 2$ in the case of a square plate, with $\varepsilon = 0.01$, $N = 10\,000$, and $\varphi_1 = \varphi_2 = 0$.

$r=\|x\|$	$K_u(r)$	$K_{M_x u}(r)$
0	1.00	0.99
0.2	0.86	0.89
0.4	0.70	0.78
0.5	0.62	0.73
0.7	0.41	0.62

4.3.4. Now we consider a vector random estimate for the system of Lamé's equations

$$\mu\Delta u(x) + (\lambda + \mu)\,\mathrm{grad}\,\mathrm{div}\,u(x) = f, \quad x \in G \subset R^3,$$
$$u|_\Gamma = g, \tag{4.48}$$

where $\Gamma = \partial G \setminus \Gamma_0$, $\int_{\Gamma_0} d\sigma(y) = 0$. Here we shall consider only the case where the estimate is constructed on the trajectories of a process of random walk over spheres. Problem (4.48) can be reduced to a system of integral equations (see [4.8]) with a matrix integral operator K and a right-hand side $\tilde{g}$, a vector which is a continuation of g into ∂G_ε. It is clear that the usual method of random walk over spheres implements a simple iteration process for that system,

i.e. is equivalent to summation of the vector series

$$u(x) = \sum_{n=0}^{\infty} \lambda^n K^n \tilde{g} \tag{4.49}$$

for $\lambda = 1$. In the general case this series is divergent, however, so we use below an analytic continuation with respect to the parameter λ and construct a uniformly and absolutely convergent series. Consider the matrix $C = \{c_{ij}\} = \{a\delta_{ij} + b\}$, where $a = 1 - \beta(\lambda, \mu)$, $b = 3\beta(\lambda, \mu)$, $\beta(\lambda, \mu) = 5(\lambda + \mu)/[2(\lambda + 4\mu)]$. Then series (4.49) is obviously majorized by the series

$$u_0 = \sum_{n=0}^{\infty} \lambda^n C^n \tilde{g} \quad (\lambda = 1). \tag{4.50}$$

Now we use an analytic continuation with respect to the parameter λ to go from this series to a new, uniformly and absolutely convergent series

$$u_0 = \sum_{n=0}^{\infty} l_n C^n \tilde{g}. \tag{4.51}$$

Since the eigenvalues of C are $1 - \beta$, $1 + 8\beta$, the substitution $\lambda = \varphi(\eta)$ can be constructed as follows:

$$\lambda = \varphi(\eta) = \frac{1}{a+b} - i\left[\frac{2a+b}{a(a+b)\pi} \ln\left(1 - \frac{z_0 - \eta\bar{z}_0}{1-\eta}\right) + \frac{b}{a(a+b)\pi} \ln\left(1 + \frac{2a+b}{b} \frac{z_0 - \eta\bar{z}_0}{1-\eta}\right)\right]$$

where z_0 is determined from the equation

$$-ia\pi = (2a+b)\ln(1-z_0) + b\ln\left(1 + \frac{2a+b}{b} z_0\right).$$

In representation (4.51)

$$l_k = \sum_{n=k}^{m} d_k^{(n)} \eta_0^n, \quad d_k^{(n)} = \frac{1}{n!}\left\{\frac{\partial^n}{\partial \eta^n}[\varphi(\eta)]^k\right\}_{\eta=0}, \quad \eta_0 = \varphi^{-1}(1).$$

It is clear that with a change of variables it is possible to construct a single mapping suitable for an analytic continuation of all the three scalar series.

So the solution of the original problem is constructed in the form

$$u = \sum_{k=1}^{n} l_k K^k g$$

and the coefficients are calculated in advance. For example, with $\varepsilon = 0.01$, $\lambda = 0.5$, $\mu = 1$, it was necessary for us to compute the

first twenty values of l_k ($m = 20$): $l_0^{(20)} = 1.0$, $l_1^{(20)} = 0.999976$, $l_2^{(20)} = 0.993152$, . . ., $l_{10}^{(20)} = 0.237118$, . . ., $l_{20}^{(20)} = 0.051874$. Thus, after the process of random walk over spheres has got into ∂G_ε the vector contribution corresponding to the estimate of $l_j K^j \tilde{g}$ is calculated at the jth step.

SOLUTION OF STOCHASTIC PROBLEMS OF RADIATIVE TRANSPORT THEORY

4.4.1. In Sec. 4.3 we discussed the application of double randomization to stochastic problems of elasticity theory. Another important class of problems whose solution by the Monte Carlo method is essentially based on the use of double randomization comprises stochastic problems of radiative transport theory. Such problems arise, for example, in the theory of photographic systems, in the theory of vision, in calculating the solar radiation passing through stochastic cloud cover, and in computing the radiation field in the ocean-atmosphere system.

Consider the integral equation

$$J(\mathbf{r}, \omega) = J_0(\mathbf{r}, \omega) + \int_0^{\xi_0} d\xi \sigma_s(\mathbf{r} - \xi\omega) \exp\left[-\int_0^{\xi} \sigma(\mathbf{r} - \xi'\omega)\, d\xi'\right] \times \frac{1}{2\pi} \int_\Omega d\omega' g(\mathbf{r} - \xi\omega, \mu) J(\mathbf{r} - \xi\omega, \omega'), \quad (4.52)$$

where

$$J_0(\mathbf{r}, \omega) = S(\mathbf{r} - \xi_0\omega, \omega) \exp\left[-\int_0^{\xi_0} \sigma(\mathbf{r} - \xi\omega)\, d\xi'\right],$$

which describes the process of radiative transport in scattering and absorbing medium in the ray optics approximation [4.1], [4.11]. Here $J(\mathbf{r}, \omega)$ is the intensity of radiation at a point $\mathbf{r}$ in the direction ω; $\sigma(\mathbf{r})$ and $\sigma_s(\mathbf{r})$ are the attenuation and scattering coefficients, respectively, $g(\mathbf{r}, \mu)$ is the scattering indicatrix with normalization $\int_{-1}^{+1} g(\mathbf{r}, \mu)\, d\mu = 1$, $\mu = \cos(\omega\omega')$, ξ_0 is the distance from a point $\mathbf{r}$ to the boundary of the medium in the direction ω; and Ω is the set of all directions. The form of the function $S(\mathbf{r}, \omega)$ determined by the source distribution and the character of the interaction between the radiation and the boundary of the medium will later be specified. Let some characteristic of the medium (for example, the

attenuation coefficient $\sigma(\mathbf{r})$ or the boundary of the medium) be a random field. We consider the problem of estimating the average field distribution value of some functional J of the solution of equation (4.52): $J = (J, \psi)$, ψ is a nonnegative function. In Sec. 4.2 we gave constructive algorithms for simulating the fields and a method for optimizing the randomization by splitting in the general case. In this section we present particular examples of using randomization of algorithms in two typical problems of radiative transport theory.

4.4.2. Consider the problem of calculating a radiation field in the ocean-atmosphere system in plane geometry. In this case the function $S(\mathbf{r}, \omega)$ in (4.52) is of the form

$$S(\mathbf{r}, \omega) = \begin{cases} \delta(\omega - \omega_0), & \omega \in \Omega_-, \quad z = H, \\ J(\mathbf{r}, \omega), & \omega \in \Omega, \quad z = h + \zeta(\mathbf{r}_\perp), \\ J(\mathbf{r}, \omega) & \omega \in \Omega_+, \quad z = 0, \end{cases} \tag{4.53}$$

where

$$J(\mathbf{r}, \omega)|_{\omega \in \Omega,\ z = h + \zeta(r_\perp)} = R_\zeta J,$$

$$J(\mathbf{r}, \omega)|_{\omega \in \Omega_+,\ z = 0} = R_0 J,$$

and ξ_0 is given as

$$\xi_0 = \begin{cases} (z - H)/c, & z \in [h + \zeta(\mathbf{r}_\perp), H], \quad \omega \in \Omega_-, \\ (z - h + \zeta(\mathbf{r}_\perp))/c, & z \in [h + \zeta(\mathbf{r}_\perp), H], \quad \omega \in \Omega_+, \\ (z - h + \zeta(\mathbf{r}_\perp))/c, & z \in [0, h + \zeta(\mathbf{r}_\perp)], \quad \omega \in \Omega_-, \\ z/c, & z \in [0, h + \zeta(\mathbf{r}_\perp)], \quad \omega \in \Omega_+. \end{cases}$$

Here the planes $z = 0$ and $z = H$ coincide with the ocean bottom and the upper boundary of the atmosphere, respectively, the plane $z = h$ coincides with the average level of the ocean surface, $\zeta(\mathbf{r}_\perp)$ is a realization of a random surface, where $\mathbf{r}_\perp$ is the projection of a radius vector $\mathbf{r}$ onto the horizontal plane, c is the direction cosine of the vector ω with respect to the z-axis, Ω_+ and Ω_- are hemispheres with $c \in [0, 1]$ and $c \in [-1, 0]$, respectively, R_0 is the linear integral operator describing the law of the reflection of radiation by the plane $z = 0$, and the operator R_ζ describes the interaction between the radiation and a random water-air interface. The radiation source is an infinitely broad radiation flux of unit power incident upon the upper boundary of the atmosphere $z = H$ in the direction ω_0. Assuming that the random interface is spatially homogeneous and stationary it is not hard to obtain an expression for the operator $R_{\zeta, m}$ which describes the mth interaction between a ray and the

interface at a point $\mathbf{r} = (\mathbf{r}_\perp, h + \zeta(\mathbf{r}_\perp))$:

$$R_{\zeta, m}J = \int_{\Omega_+} dn \int_{\Omega} p_m(\eta, \mathbf{n}) R(\omega', \mathbf{n}) J(\mathbf{r}_\perp, h+\eta, \omega') \times \delta[\omega' + \omega_{\text{refl}}(\omega, \mathbf{n})] d\omega' + \nu^2 \int_{\Omega_+} dn \int_{\Omega} p_m(\eta, \mathbf{n}) \times [1 - R(\omega', \mathbf{n})] J(\mathbf{r}_\perp, h+\eta, \omega') \delta[\omega' - \omega_{\text{refr}}(\omega, \mathbf{n})] d\omega'. \quad (4.54)$$

Here $p_m(\eta, \mathbf{n}) = p_m(\eta, \mathbf{n}/\varphi_{m-1}, e_{m-1}, \Delta x_{m-1}, \Delta y_{m-1}; \ldots; \zeta_1, e_1, \Delta x_1, \Delta y_1)$ is the conditional probability density of distribution of the deviations of the surface η from the average level h and the normal to it $\mathbf{n}$ at the point $(\mathbf{r}_\perp, h + \eta)$ of the mth interaction ($m = 2, 3, \ldots$) between the ray and the surface provided at the points of the previous interactions $(\mathbf{r}_\perp^{(m-1)}, h + \eta_{m-1}), \ldots, (\mathbf{r}_\perp^{(1)}, h + \eta_1)$ the deviations and normals were $\eta_{m-1} = \zeta_{m-1}, \ldots, \eta_1 = \zeta_1$ and $\mathbf{n}_{m-1} = \mathbf{e}_{m-1}, \ldots, \mathbf{n}_1 = \mathbf{e}_1$ respectively, Δx_i and Δy_i are the distances between the points $\mathbf{r}_\perp$ and $\mathbf{r}_\perp^{(i)}$ along the x- and y-axes, $i = \overline{1, m-1}$, respectively, the unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_{m-1} \in \Omega_+$, and $p_1(\eta, \mathbf{n})$ is the unconditional probability density of the distribution of η and $\mathbf{n}$. In (4.54) the reflection coefficient $R(\omega, \mathbf{n})$ and the unit vectors of the reflection and refraction directions, $\omega_{\text{refl}}(\omega, \mathbf{n})$ and $\omega_{\text{refr}}(\omega, \mathbf{n})$, are calculated from the familiar Fresnel formulas (see [4.1], [4.11], for example),

$$\nu = \begin{cases} n, & (\omega'\mathbf{n}) \leqslant 0; \\ 1/n, & (\omega'\mathbf{n}) > 0, \end{cases}$$

n is the index of refraction of water with respect to air. Expression (4.54) corresponds with the ocean surface interaction model in which the reflection and refraction of rays occur on an elementary area which is tangent to the surface at the point of interaction. In the geophysical literature this model is sometimes called a facet model.

Problem (4.52) to (4.54) can be solved by double randomization according to the following scheme:

(i) given a distribution density of deviations simulate in some way a realization of a random surface $z(\mathbf{r}_\perp) = h + \zeta(\mathbf{r}_\perp)$;

(ii) for the surface realization obtained simulate several trajectories (for details of constructing trajectories see [4.1]) on which the functional $J(\zeta)$ is estimated [4.11];

(iii) estimate the desired quantity $\langle J(\zeta) \rangle$ by statistical averaging of the values of $J(\zeta)$ over the ensemble of all the realizations of the random surface.

The optimum number of trajectories to be simulated for a single realization of the surface can be obtained from formulas (4.18) and (4.19). An important, and the most difficult, question arising

in solving the problem is that of constructing realizations of a random surface. As will be seen in what follows, in practice the solution of this question is closely related to the form of the desired functional.

The data on statistical properties of wind waves of the ocean surface indicate that the waves can be described with a very high accuracy by a Gaussian random field of deviations from the average level (see [4.12], for example). As is well known, such a field is fully determined by a correlation function. Hence, to construct a realization of a random surface one can use either of the methods of simulating the random fields described in Sec. 4.2 (recall that the first method is suitable for simulating a field only when the correlation function is convex). The main disadvantage of the first method is that it is necessary to use smoothing splines to construct a smooth surface from the field realization obtained. Choosing a suitable three-dimensional spline is not a trivial matter, therefore the second method of simulating according to formula (4.24) is more preferable.

There is a practically important class of problems which involves observations of the ocean from heights well in excess of typical amplitudes of the waves. In such problems the surface deviations from the average level can be ignored. In that case the interface can be described by a random surface made up of a collection of elementary areas whose centres are in the plane $z = h$, with the normal to the elementary area distributed according to the normal law with a given correlation function. This formulation of the simulation of realization of the surface implies that at the mth interaction between a ray and the plane $z = h$ a random direction of the normal $\mathbf{n}$ to the surface is chosen at the point of interaction in accordance with the conditional distribution density of the normals

$$p_m(\mathbf{n}) = p_m(n/e_{m-1}, \Delta x_{m-1}, \Delta y_{m-1}; \ldots; e_1, \Delta x_1, \Delta y_1). \quad (4.55)$$

Note that in geophysical calculations the representation for density (4.55) is often used in the form $p_m(n) = p_1(n)$. The range of applicability of such a representation, which corresponds with the "white noise" model, to the distribution of the actual sea surface is very limited. However, in the special case of the problem the probability of a ray of light intersecting with the surface at close points lying at a distance comparable with the slope correlation radius is very small because of a strong anisotropy of scattering and relatively great free length of the ray in the water. Therefore such a simple model, as numerical estimates show, gives in a number of cases a satisfactory picture of the influence of a random surface on the desired functional.

In conclusion, to illustrate the influence of a stochastic surface on the value of $\langle J \rangle$ Fig. 4.1 presents the results of angular distribution calculations of the brightness of the ocean-atmosphere system

on the upper boundary of the atmosphere in the region of a patch of sunlight for two values of the wind velocity, $v = 2$ and 5 m/s. The results have been obtained for the zenith distance of the Sun of 60°.

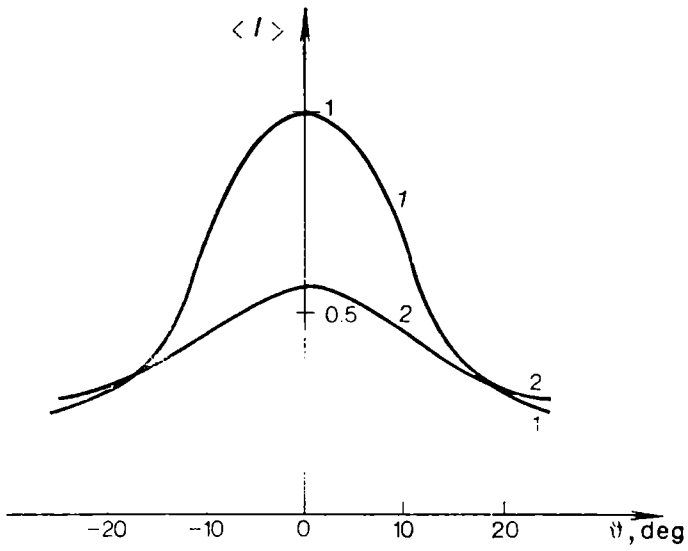


Fig. 4.1. The brightness of the ocean-atmosphere system in the region of a patch of sunlight. *1*: $v = 2$ m/s, *2*: $v = 5$ m/s.

The calculations assumed the optical ocean and atmosphere model for a sunlight wavelength of 0.5 μm described in [4.11]. The undulatory surface model was taken to be a "white noise" model with the slope distribution density

$$p_1(z'_x, z'_y) = [1/(2\pi\alpha_x\alpha_y)] \exp\{-(z'_x/\alpha_x)^2 - (z'_y/\alpha_y)^2\}$$

for the following parameters: $\alpha_x = 7.95 \times 10^{-2}$, $\alpha_y = 8.27 \times 10^{-2}$ at $v = 2$m/s and $\alpha_x = 1.26 \times 10^{-1}$, $\alpha_y = 1.12 \times 10^{-1}$ at $v = 5$ m/s. The results show that an increase in the wind velocity from 2 to 5 m/s results in an almost two-fold decrease in the brightness of the patch of sunlight.

4.4.3. As another example of the use of randomization we consider the problem of estimating the probability p of radiation passing through a slab $0 \leqslant z \leqslant H$ of material whose attenuation coefficient is a random field $\sigma(z)$. In this case in (4.52)

$$S(r, \omega) = S(z, \omega) = \begin{cases} \delta(\omega - \omega_0) \cdot & \omega \in \Omega_-, \quad z = H, \\ 0. & \omega \in \Omega_+. \quad z = 0, \end{cases}$$

$$p = \langle p(\sigma) = (J(\sigma), \psi) \rangle,$$

where

$$\psi = \begin{cases} 1, & z \leqslant 0, \quad \omega \in \Omega_-, \\ 0 & \text{otherwise.} \end{cases}$$

For a determinate slab there is a well-known method of optimizing the calculation of p on the basis of a multiple splitting technique [4.4], [4.13]. Such optimization is reduced to finding the number of splitting boundaries k^0 and splitting parameters l_i^0 $(i = \overline{1, k^0})$ minimizing the cost of the algorithm. In the stochastic case the problem arises of a simultaneous determination of optimum values of $k^0(\sigma)$, $l_i^0(\sigma)$ $(i = \overline{1, k^0(\sigma)})$ and $n^0(\sigma)$, the number of conditionally independent "combined" trajectories which are constructed for a single realization of σ. By "combined" trajectory we mean a collection of conditionally independent trajectories obtained by k-fold splitting of the "original" trajectory. Consider incorporating the choice of $n^0(\sigma)$ into the scheme of multiple splitting. The unbiased random estimate of the functional p in the case where for a single realization of σ $n(\sigma)$ "combined" trajectories are constructed from a collection of splitting parameters $\{l_1(\sigma), \ldots, l_k(\sigma)\}$ is of the form

$$\eta_{n, l_1, \ldots, l_k}(\sigma) = \frac{1}{n(\sigma) l_1(\sigma) \ldots l_k(\sigma)} \sum_{i=1}^{n(\sigma)} \sum_{j_1=1}^{l_1(\sigma)} \ldots \sum_{j_k=1}^{l_k(\sigma)} \psi(\xi_1^{(i)}, \xi_2^{(j_1, i)}, \ldots, \xi_{k+1}^{(j_k, i)}).$$

Here the random variables $\xi_m^{(1, i)}, \ldots, \xi_m^{(l_{m-1}(\sigma), i)}$ $(m = \overline{2, k+1})$ are independent and equally conditionally distributed provided $\xi_{m-1} = z_{m-1}$, z_{m-1} being the coordinate of the $(m-1)$st plane of the splitting. The variance of this estimate and the average time required to calculate a single sampling value of η are defined by the expressions

$$D\eta_{n, l_1, \ldots, l_k}(\sigma) = D_\sigma M_{\xi_1, \ldots, \xi_{k+1}}(\eta/\sigma) + \left\langle \frac{D_{\xi_1, \ldots, \xi_{k+1}}(\eta/\sigma)}{n(\sigma)} \right\rangle, \tag{4.56}$$

$$T\eta_{n, l_1, \ldots, l_k}(\sigma) = T_0 + \langle n(\sigma) \{T_1(\sigma) + \sum_{m=1}^{k} M_{\xi_1, \ldots, \xi_m}[l_1 \ldots l_m T_{m+1}(\xi_m/\sigma)]\} \rangle.$$

Here T_0 is the average time required to simulate a realization of $\sigma(z)$, $T_m(\xi_{m-1}/\sigma)$ is the average time required to calculate a single sampling value of ξ_m provided $\xi_{m-1} = z_{m-1}$, $m = \overline{1, k+1}$, and $D_{\xi_1, \ldots, \xi_{k+1}}(\eta/\sigma)$ are determined as in [4.13]. Note that in this formulation it is rather hard to obtain a constructive algorithm for estimating optimum values of $k^0(\sigma)$, $l_1^0(\sigma)$, $\ldots$, $l_{k^0}^0(\sigma)$, $n^0(\sigma)$. Putting $k(\sigma) = k$, $n(\sigma) = n$, $l_1(\sigma) = l_1$, $\ldots$, $l_k(\sigma) = l_k$ we have

for (4.56)

$$D\eta_{n, l_1, \ldots, l_k}(\sigma) = D_0 + (1/n)\left[D_1 + \sum_{m=1}^{k} (D_{m+1}) / \prod_{j=1}^{m} l_j\right],$$

$$T\eta_{n, l_1, \ldots, l_k}(\sigma) = T_0 + n\left(T_1 + \sum_{m=1}^{k} T_{m+1} \prod_{j=1}^{m} l_j\right),$$

where $D_\sigma = D_\sigma M_{\xi_1, \ldots, \xi_{k+1}}(\eta/\sigma)$ and D_m are determined as in [4.13]. It is known [4.4], [4.13] that optimum splitting parameters are calculated from the formulas

$$n^0 \approx \sqrt{D_1 T_0 / D_0 T_1}, \quad l_m^0 \approx \sqrt{D_{m+1} T_m / D_m T_{m+1}}, \quad m = \overline{1, k}. \tag{4.57}$$

The methods of estimating D_m and T_m ($m = \overline{1, k+1}$) were discussed in [4.4], [4.13]. Thus the choice of optimum splitting parameters can be made according to the following scheme.

1. Make a crude estimate of p. Use the relation (see [4.13])

$$\frac{k}{k+1} \frac{1 - p^{-1/k}}{1 - p^{-1/(k+1)}} = 1 \tag{4.58}$$

to find an appropriate value of k^0. An analysis of (4.58) shows that if the determination of p is in error by one order, then the estimation of k is in error by the unity only.

2. In accordance with the recommendations in [4.4], [4.13] form some collection of splitting parameters for a given k^0 and estimate the values of D_m and T_m ($m = \overline{0, k+1}$).

3. Evaluate n^0, l_1^0, . . ., l_{k0}^0 by substituting the obtained values of D_m and T_m into (4.57).

To illustrate the applicability of the above algorithm we give a numerical example. A homogeneous layer $0 \leqslant z \leqslant 1$ was considered upon which radiation was incident in the direction reverse to the z-axis. The following characteristics of the medium were assumed: $\sigma_s/\sigma = 0.7$, $g(\mu) = 0.05 + 0.9\delta(\mu - 1)$ and the attenuation coefficient σ was distributed with density $f_\sigma(x) = \exp\{(s - x)/t\}/t$, $s \leqslant \sigma \leqslant \infty$, $s = 20$, $t = 40$. The value of the desired probability in this case is $p = 0.4 \times 10^{-4}$. According to (4.58) $k^0 = 5$. Given a homogeneous chain of splitting [4.4], [4.13] for which $l_1 = \ldots = l_k = l$ the algorithm gave $n^0 = 4$ and $l^0 = 4$. Numerical experiments with variations in the values of n and l have shown that the actual minimum cost of the algorithm is attained with $n^0 = 4$ and $l^0 = 3$ and is about 1.3 times less in value than the cost obtained when $n^0 = l^0 = 4$. The splitting parameters obtained from formulas (4.57) are thus in good agreement with the results of the numerical experiments. Note that on the whole the above optimization of the algorithm for this example has given a more than twenty-fold gain

in the efficiency of computations as compared to an unoptimized method of randomization.

Further an example is given which illustrates the effect of a medium stochasticity on the functional p. In the layer $0 \leqslant z \leqslant H$ a stationary random field of the attenuation coefficient $\sigma(z)$ is given.

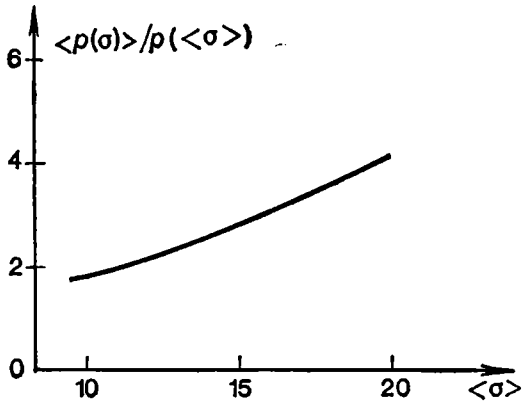


Fig. 4.2. The dependence of the ratio $\langle p(\sigma)\rangle/p(\langle\sigma\rangle)$ on $\langle\sigma\rangle$.

The field is defined by a one-dimensional distribution density $f_\sigma(x) = \exp(-x/t)/t$ and a correlation function $K(u) = \exp(-tu)$, $u = |z_1 - z_2|$, $0 \leqslant z_1, z_2 \leqslant H$. Realizations of the random field were constructed using the Palm fluxes (see Sec. 4.2). In this case for the densities $f(u)$ and $f_1(u)$ (in the notation of Sec. 4.2) we have $f(u) = f_1(u) = te^{-tu}$ and the independent random variables $\{\eta_i\}$ are calculated from the formulas $\eta_i = -(\ln \alpha_i)/t$, where $\{\alpha_i\}$ are independent random numbers uniformly distributed in the interval $[0, 1]$. For every value of η_i the value of σ is calculated from the formula $\sigma = -t \ln \alpha$. Figure 4.2 gives the dependence of the ratio $\langle p(\sigma)\rangle/p(\langle\sigma\rangle)$ on the value of $\langle\sigma\rangle$ for the following parameters of the medium: $H = 1$, $\sigma_s/\sigma = 0.7$, $g(\mu) = 0.05 + 0.9\delta(\mu - 1)$. This example shows that the transmission of a stochastic layer exceeds that of a determinite layer with an average optical thickness, this difference increasing with an increase in the average optimal thickness.

4.5. ALGORITHMS OF THE MONTE CARLO METHOD FOR SOLVING RADIATIVE TRANSPORT PROBLEMS WITH POLARIZATION

4.5.1. The scalar transport equation (see [4.1], for example) describes sufficiently well radiative transfer processes in material and is widely used to calculate various functionals expressing energetic characteristics of radiation multiply scattered by a

medium: the amount of radiation absorbed and transmitted or reflected by a layer, spatial and angular distribution of radiation intensity, etc. Numerical experiments carried out to study what errors are made in calculations of these quantities because of neglecting polarization effects show that it is possible to solve many practical and theoretical problems of the optics of planetary atmospheres in the approximation of the scalar transport equation [4.1]. Errors arising in the calculation of radiation intensity, for example, amount to no more than 10% in the most unfavourable cases. It is clear, however, that in solving radiative transfer problems requiring great accuracy, inverse problems for example, it is necessary to take the polarization of radiation into consideration. Moreover, the polarization characteristics, such as the degree of polarization and the position of neutral points, carry themselves much information about optical properties of a scattering medium and can be used in formulating and solving inverse problems of atmospheric optics on the basis of radiation and polarization measurements.

Atmospheric radiative transfer with polarization is described by a system of integro-differential transfer equations in the vector function of light intensity $\mathbf{I}(\mathbf{r}, \boldsymbol{\omega}) = (I_1, I_2, I_3, I_4)'(\mathbf{r}, \boldsymbol{\omega})$ which was first obtained and analyzed by Chandrasekhar [4.14] and Rozenberg [4.15]. For the Monte Carlo method it is convenient to consider the corresponding system of integral equations in the vector function of collision density $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)'$ related to the vector function $\mathbf{I}$ by

$$\boldsymbol{\varphi}(\mathbf{r}, \boldsymbol{\omega}) = \mathbf{I}(\mathbf{r}, \boldsymbol{\omega})\, \sigma(\mathbf{r}). \tag{4.59}$$

We denote by $\mathbf{f} = (f_1, f_2, f_3, f_4)'$ the vector function of the distribution density of a source and write the integral transfer equation with polarization:

$$\boldsymbol{\varphi}(\mathbf{x}) = \int_X K(\mathbf{x}', \mathbf{x})\, \boldsymbol{\varphi}(\mathbf{x}')\, d\mathbf{x}' + \mathbf{f}(\mathbf{x})$$

or

$$\varphi_i(\mathbf{x}) = \sum_{j=1}^{4} \int_X k_{ij}(\mathbf{x}', \mathbf{x})\, \varphi_j(\mathbf{x}')\, d\mathbf{x}' + f_i(\mathbf{x}). \tag{4.60}$$

Here $\mathbf{x} = (\mathbf{r}, \boldsymbol{\omega})$ is a point of the phase space $X = R \times \Omega$ of coordinates $\mathbf{r} \in R$ and directions $\boldsymbol{\omega} \in \Omega$, with $\mathbf{r} = (x, y, z)$ and $\boldsymbol{\omega} = (a, b, c)$. The transition matrix $K(\mathbf{x}', \mathbf{x})$ is of the form

$$K(\mathbf{x}', \mathbf{x}) = \frac{\sigma_s(\mathbf{r}') \exp(-\tau(\mathbf{r}', \mathbf{r}))\, \sigma(\mathbf{r})}{\sigma(\mathbf{r}')\, 2\pi\, |\mathbf{r}-\mathbf{r}'|^2}\, \delta\left(\boldsymbol{\omega}' - \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|}\right) P(\boldsymbol{\omega}', \boldsymbol{\omega}, \mathbf{r})$$

$$= \{k_{ij}(\mathbf{x}', \mathbf{x})\}, \quad i, j = 1, 2, 3, 4. \tag{4.61}$$

Here $P(\omega', \omega, \mathbf{r})$ is the angular matrix with elements $p_{ij}(\omega', \omega, \mathbf{r})$, $i, j = 1, 2, 3, 4$, related to the scattering matrix $R(\mu, \mathbf{r})$ as $P(\omega', \omega, \mathbf{r}) = L(\pi - i_2) R(\mu, \mathbf{r}) L(-i_1)$, where $L(i)$ is the rotation matrix (see [4.14]), i_1 and i_2 are the angles between the scattering plane and the planes through the axis of the coordinate system and the vectors ω' and ω, respectively, $\mu = \omega' \cdot \omega$ is the cosine of the scattering angle, $\sigma(\mathbf{r})$ is the total attenuation coefficient, σ_s is the scattering coefficient, and $\tau(\mathbf{r}', \mathbf{r})$ is the optical length of the path from $\mathbf{r}'$ to $\mathbf{r}$. In what follows it is assumed that the components of the vector $\mathbf{I}(\mathbf{r}, \omega)$ coincide with the Stokes parameters I, Q, U, V. Then the first element of the matrix $R(\mu, \mathbf{r})$ is the scattering indicatrix in the ordinary sense and satisfies the normalization condition

$$\int_{-1}^{1} R_{11}(\mu, \mathbf{r})\, d\mu = 1$$

In the general case of an anisotropic medium the matrix R contains all 16 elements, but the number of independent components diminishes with different properties of scattering medium and particles. In particular, in the case of an isotropic medium $\mathbf{R}$ contains only 8 independent elements (R_{11}, R_{12}, R_{21}, R_{22}, R_{33}, R_{34}, R_{43}, R_{44}) and the remaining elements are zero. Such block matrices are used to describe the laws of scattering of the Mie type [4.16]. If the scattering particles are homogeneous spheres, then $R_{11} = R_{22}$, $R_{12} = R_{21}$, $R_{33} = R_{44}$ and $R_{34} = -R_{43}$. Molecular scattering is described sufficiently well by a matrix of precisely this form, where $R_{11} = R_{22} = (3/8)(1 + \mu^2)$, $R_{21} = R_{12} = (3/8)(\mu^2 - 1)$, $R_{33} = R_{44} = (3/4)\mu$, $R_{34} = R_{43} = 0$. After scattering at a point $\mathbf{r}$ the Stokes vector parameter $\mathbf{I}'$ is transformed by the matrix P into the vector parameter $\mathbf{I}$ as follows [4.14]:

$$\mathbf{I}(\mathbf{r}, \omega) = \mathbf{P}(\omega', \omega, \mathbf{r})\, \mathbf{I}'(\mathbf{r}, \omega). \tag{4.62}$$

The existence of a solution of the system of integral transfer equations (4.60) is stipulated by the convergence of the corresponding Neumann series. We write system (4.60) as

$$\varphi_i(\mathbf{x}) = [\mathbf{K}\varphi]_i(\mathbf{x}) + f_i(\mathbf{x}), \quad i = 1, 2, 3, 4. \tag{4.63}$$

We shall consider it in the space L_1; $\varphi, \mathbf{f} \in L_1$, $\mathbf{K} \in [L_1 \to L_1]$, with

$$\| \varphi \| = \sum_{i=1}^{4} \int_X |\varphi_i(\mathbf{x})|\, d\mathbf{x},$$

$$\| \mathbf{K} \| = \sup_{j, x'} \sum_{i=1}^{4} \int_X |k_{ij}(\mathbf{x}', \mathbf{x})|\, d\mathbf{x}.$$

It is known that if the norm of the operator $\mathbf{K}$ or of some of its powers n_0 is less than unity, then the solution of system (4.63) can be represented as a Neumann series

$$\varphi = \sum_{n=0}^{\infty} \mathbf{K}^n \mathbf{f}, \quad \mathbf{K}^0 \mathbf{f} = f.$$

Simple physical considerations that radiation intensity rapidly decreases with multiplicity shows that in light polarization problems for a cloudless spherical atmosphere the relation $\| \mathbf{K}^{n_0} \| < 1$ holds for moderate values of n_0, i.e. the spectral radius $\rho(\mathbf{K}) < 1$. A numerical analysis of the behaviour of $\| \mathbf{K} \|$ is a rather difficult task. The following inequality (see 4.61) is true, however,

$$\| \mathbf{K}^{n_0} \| \leqslant \| \mathbf{K}_{sc}^{n_0} \| \sup_{\omega', \mathbf{r}} [p_{11}(\omega', \omega, \mathbf{r})/G(\omega', \omega, \mathbf{r})]^{n_0}$$

$$= \| \mathbf{K}_{sc}^{n_0} \| \cdot F_{n_0}, \quad (4.64)$$

where $\mathbf{K}_{sc}$ is the integral operator for the scalar transport equation and $G(\omega', \omega, \mathbf{r})$ is the phase function. (In (4.64) the following property of the elements of the scattering matrix is used [4.16]: $| p_{ij} | \leqslant p_{11}$, $i, j = 1, 2, 3, 4$.) It follows that one can gain an insight into the behaviour of $\| \mathbf{K}^{n_0} \|$ from the function F_{n_0} and that the spectral radius $\rho(\mathbf{K}) < 1$ if $\lim_{n \to \infty} F_n = \text{const}$. Numerical experiments designed to study the behaviour of F_n were carried out for a plane-parallel atmosphere nonhomogeneous in height. The optical thickness of the atmosphere ranged from 0.1 to 20 and the albedo of the Earth's surface was 0, 0.25, and 0.8. Two cases were studied: a Rayleigh atmosphere and an actual atmosphere with both molecular and aerosol scattering. The aerosol scattering matrix was calculated from the Mie theory. The results of the calculations will be presented in Sec. 4.5.2.

4.5.2. It is hardly possible to prove for the system of transfer equations that the spectral radius of the operator $\mathbf{K}_1$ obtained from $\mathbf{K}$ by the substitution $k_{ij} \to | k_{ij} |$, $i, j = 1, 2, 3, 4$, is less than unity (see Sec. 4.1). It is possible, however, to use some properties of the Stokes vector parameter $\mathbf{I} = (I_1, I_2, I_3, I_4)$ to prove the applicability of the algorithms of the Monte Carlo method.

It is shown in [4.1] that $\mathbf{K}$ is the compression operator on the set F of Stokes vector functions $\varphi = (I_1, I_2, I_3, I_4) = (I_1[\varphi], I_2[\varphi], I_3[\varphi], I_4[\varphi])$. It is known that the components of the Stokes vector have the following properties (see [4.14]):

$$I_1 \geqslant 0, \quad I_2^2 + I_3^2 + I_4^2 \leqslant I_1^2. \quad (4.65)$$

It follows that

$$|I_2| + |I_3| + |I_4| \leqslant \sqrt{3} I_1$$

and

$$\| \varphi \| \leqslant (1+\sqrt{3}) \int_X I_1 \, d\mathbf{x} = (1+\sqrt{3}) \| I_1 [\varphi] \|. \tag{4.66}$$

The operator $\mathbf{K}$ can be represented as $\mathbf{K} = D \cdot S$, where S is the scattering operator and D is the attenuation operator. It follows immediately from the physical meaning of these operators that $\mathbf{K}$, D, $S \in [F \to F]$; $\| I_1 [S\varphi] \| = \| I_1 [\varphi] \|$, i.e. the integrated intensity is preserved after scattering, and that $\| I_1 [D\varphi] \| \leqslant q \| I_1 [\varphi] \|$, where $q < 1$ for a medium of finite dimensions or in the presence of absorption. Combining the last relations we obtain the inequality

$$\| \mathbf{K}^n f \| \leqslant (1+\sqrt{3}) \| I_1 [\mathbf{K}^n f] \| \leqslant (1+\sqrt{3}) \| I_1 [f] \| q^n$$

which proves the existence of a solution of the system of integral transfer equations in the class of Stokes vector functions.

The main characteristics of the radiative transfer process are expressed by linear functionals of the form

$$J_H^{(i)} = (\mathbf{H}^{(i)}, \varphi) = \int_X \mathbf{H}^{(i)}(\mathbf{x}) \varphi(\mathbf{x}) \, d\mathbf{x}$$

of a solution φ of system (4.60), where $\mathbf{H}^{(i)} = \{h_k^{(i)}(\mathbf{x})\}$, $i, k = 1, \ldots, 4$, are vectors, with positive bounded components, determined by the character of the functionals to be calculated. Taking into account the representation of φ as a Neumann series we have

$$J_H^{(i)} = (\mathbf{H}^{(i)}, \varphi) = \sum_{n=0}^{\infty} (\mathbf{H}^{(i)}, \mathbf{K}^n \mathbf{f})$$

or, setting $\mathbf{K}^n \mathbf{f} = \boldsymbol{\gamma}^{(n)}$, $\boldsymbol{\gamma}^{(n)} = (\gamma_1^{(n)}, \ldots, \gamma_4^{(n)})$,

$$J_H^{(i)} = \sum_{n=0}^{\infty} \int_X \sum_{j=1}^{4} \gamma_j^{(n)}(\mathbf{x}) h_j^{(i)}(\mathbf{x}) \, d\mathbf{x}. \tag{4.67}$$

For series (4.67) to converge and for the Monte Carlo method to be applied it suffices to require that $\| \mathbf{K}^{n_0} \| < 1$ for some n_0 or to find a convergent series majorizing (4.67).

We define, as in Sec. 4.1, in the phase space X a homogeneous Markov chain by the functions $r_0(\mathbf{x})$ and $r(\mathbf{x}', \mathbf{x})$ and the probability $g(\mathbf{x})$ of a trajectory discontinuing in passing from a state $\mathbf{x}$. We also introduce random vector weights

$$Q_0^{(i)} = \frac{f_i(\mathbf{x}_0)}{r_0(\mathbf{x}_0)}, \qquad Q_n^{(i)} = \sum_{j=1}^{4} Q_{n-1}^{(j)} \frac{k_{ij}(\mathbf{x}_{n-1}, \mathbf{x}_n)}{r(\mathbf{x}_{n-1}, \mathbf{x}_n)} \frac{1}{1-g(\mathbf{x}_{n-1})}, \tag{4.68}$$

and consider a random vector $\xi = \{\xi_k\}$, $k = 1, \ldots, 4$,

$$\xi_k = \sum_{n=0}^{N} \sum_{j=1}^{4} Q_n^{(j)} h_j^{(k)}(\mathbf{x}_n) = \sum_{n=0}^{N} \zeta_k(\mathbf{x}_n), \tag{4.69}$$

where N is a random index of the last state of the Markov chain. If it is required to calculate by the Monte Carlo method only the first component of the Stokes vector $I_H^{(1)}$, i.e. the radiation intensity, then according to (4.65) $\zeta_1(\mathbf{x}_n)$ in the last sum is positive at every collision n and it is possible to use the general theory of the Monte Carlo method to calculate sums of the form (4.69) (see [4.4], p. 151, for example).

Thus we have

$$J_{(H)}^{(1)} = M\xi_1.$$

On the other hand, it may be observed using properties (4.65) and (4.66) of the Stokes vector function that for any n we have

$$|\zeta_k(\mathbf{x}_n)| \leqslant \sqrt{3}\,\zeta_1(\mathbf{x}_n), \quad k = 2, 3, 4.$$

Relations (4.67) and (4.69) and the additive property of the Stokes vector function (see [4.14]) justify the application of the Monte Carlo method to the solution of transfer problems with polarized radiation.

Note that as compared to Sec. 4.1 we consider here a dual form of the functional since the quantity to be calculated is a functional of the basic vector transfer equation while equation (4.1) is similar to the adjoint transfer equation. A special case of the functional in question is the solution of the adjoint equation at a given point and an arbitrary functional can be obtained by averaging such a solution. Therefore the study in Sec. 4.1 of variance of vector estimates is automatically extended to include the estimates discussed in this section.

An important question in calculations of polarization is an optimum choice of scalar density of transition probabilities $r(\mathbf{x}', \mathbf{x})$. Chosen in [4.1] for this purpose was the function $k_{11}(\mathbf{x}', \mathbf{x})$ which corresponds to radiation transfer without polarization. In simulating such a process the vector weight after scattering at a point $\mathbf{r}$ is transformed by a matrix with elements $p_{ij}(\omega', \omega, \mathbf{r})/R_{11}(\mu, \mathbf{r})$. The choice of function $k_{11}(\mathbf{x}', \mathbf{x})$ was governed by physical considerations. Besides, such a choice helped to estimate with sufficient accuracy the small errors in intensity calculations caused by ignoring polarization. It is interesting to note at this point that relative mean-square errors of the results obtained in calculating the intensity without polarization are insignificantly smaller than wih polarization. For values of the azimuth angle of observation $0 < \varphi \leqslant 90°$ the values of intensity calculated without polarization are greater,

and for angles $90^\circ < \varphi \leqslant 180^\circ$ less than the values taking into account polarization. These results have been obtained over the same trajectories for several zenith angles of observation with the height of the Sun of 53°.

The results of Sec. 4.1.3 show that to construct $r(\mathbf{x}', \mathbf{x})$ it is appropriate to use a uniformly large function from the functions $|k_{ij}(\mathbf{x}', \mathbf{x})|$ or the first eigenvalue of the matrix $K(\mathbf{x}', \mathbf{x})$. The largest eigenvalue of the matrix of Rayleigh scattering is 3/4. It is appropriate therefore to take as $r(\mathbf{x}', \mathbf{x})$ the function $k_{11}(\mathbf{x}', \mathbf{x})$ in which the scattering indicatrix is spherical. The weight Q is transformed in this case by a matrix with elements $p_{ij}(\boldsymbol{\omega}', \boldsymbol{\omega}, \mathbf{r}) \cdot 2$. Numerical experiments show that relative mean-square errors in calculations of intensity and particularly the second and third components of the Stokes vector when using such a density tend to be smaller than when choosing as $r(\mathbf{x}', \mathbf{x})$ the component $k_{11}(\mathbf{x}', \mathbf{x})$. A series of calculations for various optical thicknesses τ from 0.1 to 20 was used to study the convergence to the constant of $q_n = \max Q_n^{(1)}$ with increasing multiplicity of scattering n. The maximum for a fixed n is taken over all the trajectories (1000 in number). The vector weight $\mathbf{Q}$ is defined by formulas of the form (4.68):

$$\mathbf{Q}_n = \prod_{k=1}^{n} [\mathbf{P}(\boldsymbol{\omega}_{k-1}, \boldsymbol{\omega}_k, \mathbf{r}_k)/G(\mu_k, \mathbf{r}_k)]\,\mathbf{Q}_0, \quad \mathbf{Q}_0 = (\pi S, 0, 0, 0),$$

$$\mu_k = \boldsymbol{\omega}_{k-1} \cdot \boldsymbol{\omega}_k,$$

Since $|Q_n^{(i)}| < Q_n^{(1)}$, $i = 2, 3, 4$, for all n we consider the maximum of the first component. Numerical experiments show that the convergence in question of q_n for a purely scattering molecular atmosphere with $G = 1/2$ is quicker than that with $G(\mu) = 3(1 + \mu^2)/8$. Similar calculations were also carried out for real atmosphere models with aerosol scattering matrices calculated from the Mie formulas. The calculations show that in practice with $n = 2$-4 the value of q_n attains the order 1 and then virtually stops increasing. The function $r(\mathbf{x}', \mathbf{x})$ in these calculations was chosen as in [4.1].

Note also that with $\lim_{n \to \infty} q_n = \text{const}$ the algorithms of the Monte Carlo method for calculations of polarization have finite dispersion where the corresponding scalar estimates have finite dispersion.

4.5.3. We discuss some of the most frequently used algorithms of the Monte Carlo method for estimating the intensity and polarization characteristics of scattered light in the atmosphere. To evaluate I_1, I_2, I_3, and I_4 it is necessary to choose the corresponding coordinate system fixed at a collision point $\mathbf{r}$. A spherical system with the r-axis is convenient for a spherical atmosphere and the z-axis for a plane-parallel atmosphere. Such a choice allows one to use the symmetry properties of the geometries involved and to

develop efficent algorithms of the Monte Carlo method for estimating Stokes parameters at a given point of the phase space $\mathbf{x}^* = (\mathbf{r}^*, \omega^*)$.

Let $f(\mathbf{x}^*) = 0$. On substituting $\mathbf{x}^*$ for $\mathbf{x}$ in (4.60) and using (4.59) and (4.61) we get

$$I_i(\mathbf{x}^*) = \int_X \frac{1}{\sigma(\mathbf{r}^*)} \sum_{j=1}^{4} [k_{ij}(\mathbf{x}', \mathbf{x}^*)\, \varphi_j(\mathbf{x}')]\, d\mathbf{x}'$$

$$= \int_X \Big[\sum_{j=1}^{4} p_{ij}(\mu^*, \mathbf{r}') \frac{\sigma_s(\mathbf{r}') \exp(-\tau(\mathbf{r}', \mathbf{r}^*))}{2\pi |\mathbf{r}' - \mathbf{r}^*|^2}$$

$$\times \delta\Big(\frac{\mathbf{r}' - \mathbf{r}^*}{|\mathbf{r}' - \mathbf{r}^*|} - \omega'\Big) I_j(\mathbf{x}')\Big]\, d\mathbf{x}' = \int_X \varphi_i^*(\mathbf{r}', \omega')\, F(\mathbf{x}')\, d\mathbf{x}',$$

where $\mathbf{r}^*$ and $\omega^* = (a^*, b^*, c^*)$ are the point and direction of observation, respectively, $I_j(\mathbf{x}')$ are the components of the Stokes vector before collision at a point $\mathbf{x}'$ that are transformed after each scattering by the matrix $K(\mathbf{x}', \mathbf{x})/r(\mathbf{x}', \mathbf{x})$ according to (4.62), $F(\mathbf{x})$ is a scalar collision density corresponding to the kernel $r(\mathbf{x}', \mathbf{x})$. (The initial value of the Stokes vector $\mathbf{I}^{(0)} = (\pi S, 0, 0, 0)$.) Thus $I_i(\mathbf{x}^*)$ is represented as a linear functional of collision density. As a rule, to estimate the last integral by the Monte Carlo method one should simulate particle trajectories on a computer from the transition density $r(\mathbf{x}', \mathbf{x})$ and average the sums of the values of φ_i^* for collisions of different orders. In the case at hand this method cannot be applied because of the δ-function in the expression for φ^*. In the case of a spherical atmosphere (see [4.1], p. 63) to get rid of the δ-function the expression in the square brackets is averaged over the angle of rotation about the symmetry axis of the Earth-atmosphere-Sun system. Let Γ be a surface formed by a ray $\mathbf{r}^* - \omega^*\lambda$ ($\lambda \geqslant 0$) when rotated about the symmetry x-axis and let $\mathbf{r}_1 = (x_1\ y_1,\ z_1)$ be the point of the intersection of a direction ω and Γ. Making a rotation about the symmetry axis sends the point $\mathbf{r}_1$ into the azimuthal observation plane, the point $\mathbf{r}_1$ turning into the point $\mathbf{r}_2$ on the ray $\mathbf{r}^* - \omega^*\lambda$ and ω going into ω_1. To estimate $I_i(\mathbf{x}^*)$ it is now necessary to average over all collisions the value of $\varphi_i^*(\mathbf{x})$ equal to zero if ω does not intersect Γ after scattering at a point $\mathbf{r}$ and to

$$\varphi_i^*(\mathbf{x}) = \sum_{j=1}^{4} p_{ij}(\mu^*, \mathbf{r}_1) I_j(\mathbf{x}) \frac{\sigma_s(\mathbf{r}_1) \exp(-\tau(\mathbf{r}, \mathbf{r}_1) - \tau(\mathbf{r}_2, \mathbf{r}^*))}{4\pi^2 |a^*| \cdot \rho_1 \cdot |(\mathbf{n}, \omega)|} \sqrt{1 - n_1^2}$$

otherwise. Here $\mu^* = (\omega', \omega^*)$, $\mathbf{n} = (n_1, n_2, n_3)$ is the normal to Γ at $\mathbf{r}_1$ and $\rho_1 = \sqrt{y_1^2 + z_1^2}$. It is appropriate to use this algorithm

in simulating observations in or near the twilight region of the atmosphere.

In calculations for the illuminated region the adjoint walk method is found to be more efficient than the above algorithm [4.1]. In this case, for $I_i(\mathbf{x}^*)$ to be estimated at a point $\mathbf{x}^* = (\mathbf{r}^*, \boldsymbol{\omega}^*)$ it is necessary to simulate trajectories with initial density $r_0(\mathbf{x}) = \delta(\mathbf{r} - \mathbf{r}^*) \times \delta(\boldsymbol{\omega} + \boldsymbol{\omega}^*)$ and to average over all collisions the value of

$$\varphi_i^*(\mathbf{x}) = [e^{-\tau_0}/(2\pi^2 R^2)] \sum_{j=1}^{4} p'_{ij}(\boldsymbol{\omega}^*, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{n-1}, \mathbf{r}_1, \ldots, \mathbf{r}_n) \times p_{j1}(\boldsymbol{\omega}^{(0)}, -\boldsymbol{\omega}_{n-1}, \mathbf{r}_n),$$

where the matrix $\mathbf{P}' = \prod_{k=1}^{n-1} [\mathbf{P}(-\boldsymbol{\omega}_k, -\boldsymbol{\omega}_{k-1}, \mathbf{r}_k)/G(\boldsymbol{\omega}_k, \boldsymbol{\omega}_{k-1}, \mathbf{r}_k)]$,

$\boldsymbol{\omega}^{(0)}$ is the direction of a "unit" flux of solar radiation, $\boldsymbol{\omega}_0 = \boldsymbol{\omega}^*$, and τ_0 is the optical length of the interval from a point $\mathbf{r}_n$ to the boundary of the atmosphere in the direction $-\boldsymbol{\omega}^{(0)}$; here $G(\boldsymbol{\omega}', \boldsymbol{\omega}, r)$ is the phase function (or scattering indicatrix) appearing in the density of transition probabilities $r(\mathbf{x}', \mathbf{x})$. With this method it is thus necessary to multiply the matrices at each scattering.

For calculations of the intensity and polarization at a point $(z^*, \boldsymbol{\omega}^*)$ in a plane-parallel atmosphere an efficient algorithm of the Monte Carlo method has been developed by averaging a local estimate for plane geometry (see [4.1], p. 49) along a path L from a collision point to the boundary of the atmosphere with a weight $\sigma(z) \exp\left(-\int_0^l \sigma(t)\, dt\right)$. The Stokes vector for a singly scattered radiation $\mathbf{I}^{(1)}(z^*, \boldsymbol{\omega}^*)$ is given explicitly:

$$I_i^{(1)}(z^*, \boldsymbol{\omega}^*) = [p_{i1}(\boldsymbol{\omega}_0, \boldsymbol{\omega}^*)/(2\pi(|c^*| + c_0))] \times \sum_{k=1}^{n} (\sigma_s^{(k)}/\sigma_k J^{(k)} \exp(-\tau_k^{(0)} - \tau_k^*). \quad (4.70)$$

Here n is the number of layers into which the atmosphere inhomogeneous in height is split, $\boldsymbol{\omega}_0 = (a_0, b_0, c_0)$ is the direction of incident radiation, $\tau_k^{(0)} = \sum_{j=k+1} (z_j - z_{j-1})\, \sigma_j/|c_0|$, with z_j the height of the boundary of the jth layer, $\tau_n^{(0)} = 0$, $\sigma_s^{(k)}$ and σ_k are the scattering and attenuation coefficients in the kth layer, $\tau_k^* = \sum_{j=1}^{k-1} (z_j - z_{j-1}) \times \sigma_j/|c^*|$, $\tau_1^* = 0$, $E_k = \exp(-(z_k - z_{k-1})\, \sigma_k/|c^*|)$, $J^{(k)} = E_k(1 - \exp[-\sigma_k (z_k - z_{k-1})(c_0 + |c^*|)/|c^* \cdot c_0|])$.

Formulas similar to (4.70) can be obtained for calculations of the characteristics of multiple scattering. Let ω be a new direction of a particle after collision at a point z, $z_{i_0-1} \leqslant z < z_{i_0}$. For definiteness, assume that $z^* = 0$, i.e. the "observation" is made from the Earth's surface. Consider the case $c < 0$. Let

$$\tau_k^{(1)} = (z - z_{i_0-1})\sigma_{i_0}/|c| + \sum_{j=k+1}^{i_0-1} \tau_j,$$

$$\tau_j = (z_j - z_{j-1})\,\sigma_j/|c|, \quad k = 1, \ldots, i_0 - 1, \quad \tau_n^{(1)} = 0, \quad \tau_1^* = 0,$$

$$\tau_k^* = \sum_{j=1}^{k} (z_j - z_{j-1})\,\sigma_j/|c^*|, \quad E_k^* = \exp(-\tau_k^*), \quad k = 2, \ldots, i_0,$$

$$E_1^* = 1, \quad E_k = \exp[-(z_k - z_{k-1})\,\sigma_k/|c^*|], \quad k = 1, \ldots, n,$$

$$E_{i_0} = \exp[-(z - z_{i_0-1})\,\sigma_{i_0}/|c^*|].$$

Assume that if the subscript in the summation symbol is greater than the superscript in value, then the sum is zero. Setting

$$J^{(k)} = E_k\,(1 - \exp[-\tau_k\,(|c^*| + c)/|c^*|])$$

we obtain for the calculation of $I_i\,(z^*,\ \omega^*)$ the estimate

$$\varphi_i^*(\mathrm{x}) = [2\pi\,(|c^*| + c)]^{-1} \sum_{j=1}^{4} p_{ij}\,(\mu^*)\,I_j\,(\mathrm{x})$$

$$\times \sum_{k=1}^{i_0} J^{(k)} \exp(-\tau_k^{(1)})\,E_k^*. \qquad (4.71)$$

Now let $c > 0$. We first consider the case of reflection at the underlying surface, $z = 0$. Then $\tau_k^{(1)} = \sum_{j=1}^{k} \tau_j$, $k = 1, \ldots, n$, $\tau_0^{(1)} = 0$,

$$\tau_k^* = \sum_{j=1}^{k-1} (z_j - z_{j-1})\,\sigma_j/|c^*|, \quad k = 2, \ldots, n, \quad \tau_1^* = 0,$$

$$J^{(k)} = 1 - \exp(-\tau_k\,(|c^*| + c)/|c^*|), \quad k = 1, \ldots, n. \qquad (4.72)$$

When estimated from an arbitrary point z the value of $J^{(i_0)}$ is calculated from the formula $J^{(i_0)} = \exp(-(z - z_{i_0-1})\,\sigma_{i_0}/|c^*|) \times [1 - \exp(-\tau_{i_0}\,(|c^*| + c)/|c^*|)]$ and the remaining values $J^{(k)}$, $i_0 < k \leqslant n$ are calculated from formula (4.72) for $\tau_{i_0-1}^{(1)} = 0$, $\tau_{i_0} = (z_{i_0} - z)\,\sigma_{i_0}/c$, $\tau_k = (z_k - z_{k-1})\,\sigma_k/c$, $\tau_k^{(1)} = \sum_{j=i_0}^{k} \tau_j$, $k = i_0, \ldots$

. . ., n. Assuming $i_0 = 1$ for reflection at the underlying surface we get for the case $c > 0$

$$\varphi_i^*(\mathbf{x}) = [2\pi(|c^*| + c)]^{-1} \sum_{j=1}^{4} p_{ij}(\mu^*) I_j(\mathbf{x}) \times \sum_{k=i_0}^{n} J^{(k)} \exp(-\tau_{k-1}^{(1)}) E_k^*. \quad (4.73)$$

To estimate the components of the Stokes vector $I_i(\mathbf{x})$ by this algorithm it is thus necessary to evaluate (4.71) or (4.73) after each simulation of a new direction ω and then carry out averaging over all collisions. It was assumed in the above formulas that the scattering matrix R is independent of z. It is easy to write them for the case where R is dependent on z. Note that in calculations E_i, τ_i^*, and E_i^* are evaluated at the beginning of a program; E_{i_0} is only calculated at each collision point. The values of $\tau_i^{(1)}$ appearing in the estimates considered are used later on in simulating every next path length and $\tau_i^{(0)}$ appearing in (4.70) are used in simulating the initial path length of other trajectories.

4.5.4. To conclude, the question of the finiteness of variance of Monte Carlo estimates in transfer theory with polarization requires further study. By reasoning similar to that of Sec. 4.5.2 one can obtain equation (4.6) without requiring that the condition $\psi_1(\mathbf{x}) < +\infty$ should hold. As a rule the sufficient condition of Sec. 4.1.3 for the finiteness of variance of estimates will simplify to

$$[\text{vrai} \sup_{x, y} (\| K(x, y) \| / p(x, y))] \rho(K) < 1.$$

This last formula may serve as a basis for numerical investigation of the finiteness of variance.

On the other hand, numerical studies show the boundedness of the values q_n (see Sec. 4.5.2) as $n \to \infty$, which ensures the boundedness of variance. It may be noted that the relation $q_n = 0(n^k)$, $k > 0$, also ensures the finiteness of variance. Of practical importance therefore is the analytical investigation of the behaviour of q_n, i.e. the upper bound of the elements of the weight matrix which is multiplied by the scattering matrix (and divided by the scalar transitional density) after each collision.

References

4.1. G. I. Marchuk, G. A. Mikhailov, M. A. Nazaraliev, B. A. Kargin, R. A. Darbinyan, and B. S. Elepov, *The Monte Carlo Method in Atmospheric Optics* (Springer-Verlag, Berlin-Heidelburg-New York, 1980).

4.2. I. M. Sobol', in: *The Monte Carlo Method in the Radiation Transfer Problem*, ed. Acad. G. I. Marchuk (Moscow: Atomizdat, 1967): 232-254 (in Russian).

4.3. G. A. Mikhailov, K. K. Sabelfeld, N. N. Chentsov, in: *Actual Problems of Applied Mathematics and Mathematical Modelling* (Novosibirsk: Nauka, 1982): 69-82 (in Russian).
4.4. S. M. Ermakov and G. A. Mikhailov, *Statistical Simulation*, (Moscow: Nauka, 1982) (in Russian).
4.5. G. A. Mikhailov, *Zh. Vychisl. Matem. i Mat. Fiz.* 23, No. 3 (1983) (in Russian).
4.6. N. N. Chentsov, *Probabil. Theory and Its Appl.* 1, No. 1: 154-161 (1956) (in Russian).
4.7. K. K. Sabelfeld, *Zh. Vychisl. Matem. i Mat. Fiz.* 23, No. 2 (1983) (in Russian).
4.8. K. K. Sabelfeld, *Dokl. Akad. Nauk SSSR* 262, No. 5 (1982) (in Russian).
4.9. B. S. Elepov, A. A. Kronberg, G. A. Mikhailov, K. K. Sabelfeld, *Solving Boundary-Value Problems by the Monte Carlo Method* (Novosibirsk: Nauka, 1980) (in Russian).
4.10. V. V. Bolotin, *Random Oscillations of Elastic Systems* (Moscow: Nauka, 1979) (in Russian).
4.11. E. O. Dzhetybaev and B. A. Kargin, in: *Actual Problems of Applied Mathematics and Mathematical Modelling* (Novosibirsk: Nauka, 1962): 83-91 (in Russian).
4.12. W. J. Pierson, *Adv. Geophys.* 2, No. 93 (1955).
4.13. B. A. Kargin and V. S. Troinikov, in: *Monte Carlo Methods in Computational Mathematics and Mathematical Physics*, Part 2 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 15-24 (in Russian).
4.14. S. Chandrasekhar, *Radiative Transfer* (Oxford: Oxford University Press, 1953).
4.15. G. V. Rozenberg, *Usp. Fiz. Nauk* 56 No. 1: 77-109 (1955) (in Russian).
4.16. D. Deirmenjian, *Electromagnetic Scattering on Spherical Polydispersions* (New York, 1969).

5 Numerical Simulation of a Large-Scale Atmospheric and Oceanic Circulation

G. I. Marchuk, V. P. Dymnikov, V. B. Zalesny, V. N. Lykosov, and V. Ya. Galin

Numerical simulation is one of the most effective tools for solving urgent problems of meteorology and oceanology, the theory of climate and its changes. Over the past three decades its vast possibilities have been demonstrated in numerous studies in this country and abroad. While the early work was concerned with a qualitatively correct description of the hydrothermodynamics of the atmosphere and ocean, subsequent work included their quantitative description. A sufficiently accurate quantitative description of the characteristics of the global circulation of the atmosphere and ocean by means of numerical models of the general circulation has naturally involved using these models to solve problems of simulating actual climate and its variation. On the whole the solution of the above problems could be broken down into the following stages:

(1) constructing a mathematical-physical model of the general circulation of the atmosphere and ocean (this process involves an understanding of the relative importance of physical processes maintaining the general circulation of the atmosphere and ocean and a formation of adequate models of those physical processes);

(2) developing mathematical means for the numerical solution of nonlinear three-dimensional equations of hydrothermodynamics (constructing a discrete model);

(3) developing a complex of programs for computer calculations (computer implementation of the model);

(4) carrying out numerical experiments;

(5) processing and analyzing computational results.

In general, these steps are closely related since a better understanding of the physics of the atmospheric and oceanic circulation automatically makes new demands on computing algorithms and their computer implementation and, conversely, formulation of mathematical models for the description of particular physical processes heavily relies on the available computing algorithms and computer resources.

In this paper we dwell on all of the five stages listed above (the first is dealt with in a more fragmentary manner since it has been discussed in detail in [5.1]-[5.4]). Chosen as a basis is the model of the general atmospheric and oceanic circulation developed at the

Computing Centre of the Siberian Division of the USSR Academy of Sciences and in the Department of Numerical Mathematics of the Academy ([5.1]-[5.3]).

5.1. STATEMENT OF THE PROBLEM OF A LARGE-SCALE DYNAMICS OF THE ATMOSPHERE AND OCEAN

5.1.1. The Atmosphere. We write the equations of a large-scale hydrothermodynamics of the atmosphere in the σ-system of coordinates proposed by Phillips [5.5] and differing from a spherical system in the choice of reduced pressure σ as the vertical coordinate:

$$\frac{du}{dt} - \left(l + \frac{u}{a}\tan\varphi\right)v + \frac{1}{a\cos\varphi}\left(\frac{\partial\Phi}{\partial\lambda} + \frac{RT}{\pi}\frac{\partial\pi}{\partial\lambda}\right) = F_u, \tag{5.1}$$

$$\frac{dv}{dt} + \left(l + \frac{u}{a}\tan\varphi\right)u + \frac{1}{a}\left(\frac{\partial\Phi}{\partial\varphi} + \frac{RT}{\pi}\frac{\partial\pi}{\partial\varphi}\right) = F_v, \tag{5.2}$$

$$\frac{\partial\Phi}{\partial\sigma} = -\frac{RT}{\sigma}, \tag{5.3}$$

$$\frac{\partial\pi}{\partial t} + \frac{1}{a\cos\varphi}\left[\frac{\partial(\pi u)}{\partial\lambda} + \frac{\partial(\pi v\cos\varphi)}{\partial\varphi}\right] + \frac{\partial(\pi\dot{\sigma})}{\partial\sigma} = 0, \tag{5.4}$$

$$\frac{dT}{dt} - \frac{RT}{c_p\pi\sigma}\left[\pi\dot{\sigma} + \sigma\left(\frac{\partial\pi}{\partial t} + \frac{u}{a\cos\varphi}\frac{\partial\pi}{\partial\lambda} + \frac{v}{a}\frac{\partial\pi}{\partial\varphi}\right)\right] = F_T + \varepsilon, \tag{5.5}$$

$$\frac{dq}{dt} = F_q - (C - E), \tag{5.6}$$

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{a\cos\varphi}\frac{\partial}{\partial\lambda} + \frac{v}{a}\frac{\partial}{\partial\varphi} + \dot{\sigma}\frac{\partial}{\partial\sigma}. \tag{5.7}$$

Here t is the time, λ the longitude, φ the latitude, $\sigma = p/\pi$ (with p the pressure and π its value at the Earth's surface), u, v and $\dot{\sigma}$ the wind velocity components in the longitude, latitude, and altitude, respectively, $\Phi = gz$ the geopotential of the constant σ-surface (with g the acceleration due to gravity and z the altitude), T the temperature (in degrees Kelvin), q the specific humidity of the air, $l = 2\omega\sin\varphi$ the Coriolis parameter (with ω the angular velocity of the Earth's rotation), a the Earth's radius, R the gas constant of the air, c_p the heat capacity at constant pressure, F_u and F_v the rates of change in the moment of momentum due to the Reynolds stresses, F_T and F_q the rates of change in the temperature and specific humidity due to small-scale diffusion and mesoscale convection, ε represents diabatic influxes of heat, and C and E condensation and evaporation, respectively.

The solution of the system (5.1)-(5.6) is sought in a region G which is a spherical slab of unit height. The boundary conditions in the

σ-coordinate are as follows:

$$\dot{\sigma} = 0 \quad \text{for} \quad \sigma = 1, \tag{5.8}$$

$$\dot{\sigma} = 0 \quad \text{for} \quad \sigma = 0 \tag{5.9}$$

and, in addition, the geopotential distribution over the Earth's surface is given:

$$\Phi_S = gz_S \quad \text{for} \quad \sigma = 1, \tag{5.10}$$

where z_S is the excess of the Earth's surface level over the sea level which is always assumed undisturbed.

The singularity of system (5.1)-(5.6) is that the number of equations it contains is a unit less than the number of unknown functions. The role of the missing equation is played by the additional boundary condition (5.10) on the lower boundary of the atmosphere.*

5.1.2. The Ocean. The following system of equations is used to describe large-scale motions in the ocean:

$$H\frac{du}{dt} - H\left(l + \frac{u}{a}\tan\varphi\right)v + \frac{1}{a\cos\rho_0}\left(H\frac{\partial p}{\partial\lambda} - \frac{\partial H}{\partial\lambda}z_1\frac{\partial p}{\partial z_1}\right) = HF_{Su}, \tag{5.11}$$

$$H\frac{dv}{dt} + H\left(l + \frac{u}{a}\tan\varphi\right)u + \frac{1}{a\rho_0}\left(H\frac{\partial p}{\partial\varphi} - \frac{\partial H}{\partial\varphi}z_1\frac{\partial p}{\partial z_1}\right) = HF_{Sv}, \tag{5.12}$$

$$\frac{\partial p}{\partial z_1} = gH\rho, \tag{5.13}$$

$$\frac{1}{a\cos\varphi}\left(\frac{\partial(Hu)}{\partial\lambda} + \frac{\partial(Hv\cos\varphi)}{\partial\varphi}\right) + \frac{\partial\hat{w}}{\partial z_1} = 0, \tag{5.14}$$

$$H\frac{dT}{dt} + \gamma_T\left[\hat{w} + z_1\left(\frac{\partial H}{\partial\lambda}u + \cos\varphi\frac{\partial H}{\partial\varphi}v\right)\right] = HF_{ST}, \tag{5.15}$$

$$H\frac{dS}{dt} + \gamma_S\left[\hat{w} + z_1\left(\frac{\partial H}{\partial\lambda}u + \cos\varphi\frac{\partial H}{\partial\varphi}v\right)\right] = HF_{SS}, \tag{5.16}$$

$$\rho = \rho(T, S), \tag{5.17}$$

$$\hat{w} = w - z_1\left(\frac{\partial H}{\partial\lambda}u + \cos\varphi\frac{\partial H}{\partial\varphi}v\right),$$

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \frac{u}{a\cos\varphi}\frac{\partial}{\partial\lambda} + \frac{v}{a}\frac{\partial}{\partial\varphi} + \hat{w}\frac{\partial}{\partial z_1}.$$

*Since π, the pressure at the Earth's surface, is the function of time and the λ- and φ-coordinates, the formal addition of the equation $\partial\pi/\partial\sigma = 0$ to system (5.1)-(5.6) equates the number of equations and desired functions and makes mathematically clear the introduction of the boundary condition (5.10).

The system of equations (5.11) to (5.17) is written in terms of deviations from a fluid state stably stratified according to the density, with the density, temperature, and salinity fields represented as follows:

$$\rho = \bar{\rho}(z) + \rho', \quad T = \bar{T}(z) + T', \quad S = \bar{S}(z) + S'$$

(the primes in (5.11) to (5.17) are omitted).

The equations of the ocean dynamics, (5.11) to (5.17), are written in spherical coordinates, with geometrical vertical coordinate $\lambda\varphi z_1$ ($z_1 = z/H$, where z is the distance from the ocean undisturbed surface and $H(\lambda, \varphi)$ is the underwater topography). The region of solution of the problem, Ω, is a spherical slab of unit height, with a certain part, which corresponds to sections of dry land, cut out.

Taken as vertical boundary conditions are the "hard cover" condition at the ocean surface

$$\hat{w} = 0 \quad \text{for} \quad z_1 = 0 \tag{5.18}$$

and the bottom contour flow condition

$$\hat{w} = 0 \quad \text{for} \quad z_1 = 1. \tag{5.19}$$

In this case the number of desired functions of system (5.11)-(5.16) coincides with the number of equations and it suffices to have two kinematic conditions in the vertical.*

5.2. INTEGRAL LAWS OF CONSERVATION

The systems of equations of the hydrothermodynamics of the atmosphere and ocean described in the preceding section have a number of integral properties that must be taken into account when constructing their numerical approximations.

5.2.1. The Atmosphere. We first establish the laws of conservation for the model of the general circulation of the atmosphere.

1. We integrate the continuity equation (5.4) over the entire volume of the atmosphere. By virtue of the boundary conditions we have

$$\frac{\partial}{\partial t} \int_G \pi \, dG = 0, \quad dG = a^2 \cos\varphi \, d\varphi \, d\lambda. \tag{5.20}$$

*In general, in both the atmospheric and the oceanic problem the right-hand sides of equations (5.1) to (5.6) and (5.11) to (5.17) describing the processes of large-scale turbulence exchange contain second derivatives with respect to z-coordinates, which requires some more appropriate boundary conditions to be stated.

The integration in (5.20) is over the entire terrestrial globe. Relation (5.20) implies that the model preserves the entire mass of the atmosphere.

2. We introduce the angular momentum

$$M = a \cos \varphi \, (u + a\omega \cos \varphi).$$

It is easy to see that equation (5.1) can be written for M as

$$\frac{dM}{di} + \frac{\partial \Phi}{\partial \lambda} + \frac{RT}{\pi} \frac{\partial \pi}{\partial \lambda} = a \cos \varphi \, F_u \tag{5.21}$$

or in divergence form as

$$\frac{\partial (\pi M)}{\partial t} + \frac{1}{a \cos \varphi} \left(\frac{\partial (u \pi M)}{\partial \lambda} + \frac{\partial (v \cos \varphi \, \pi M)}{\partial \varphi} \right) + \frac{\partial (\dot{\sigma} \pi M)}{\partial \sigma} + \pi \left(\frac{\partial \Phi}{\partial \lambda} + \frac{RT}{\pi} \frac{\partial \pi}{\partial \lambda} \right) = a \cos \varphi \, \pi F_u. \tag{5.22}$$

Using (5.3) we reduce (5.22) to the form

$$\frac{\partial (\pi M)}{\partial t} + \frac{1}{a \cos \varphi} \left(\frac{\partial (u \pi M)}{\partial \lambda} + \frac{\partial (v \cos \varphi \, \pi M)}{\partial \varphi} \right) + \frac{\partial (\dot{\sigma} \pi M)}{\partial \sigma} + \frac{\partial (\pi \Phi)}{\partial \lambda} - \frac{\partial \pi}{\partial \lambda} \frac{\partial (\Phi \sigma)}{\partial \sigma} = a \cos \varphi \, \pi F_u. \tag{5.23}$$

We now integrate (5.23) over the entire volume of the atmosphere and use the boundary conditions to obtain the following law of the conservation of angular momentum:

$$\frac{\partial}{\partial t} \int_0^1 \int_G \pi M \, dG \, d\sigma = \int_G \Phi_S \frac{\partial \pi}{\partial \lambda} \, dG + \int_0^1 \int_G a \cos \varphi \, \pi F_u \, dG \, d\sigma. \tag{5.24}$$

Note that since λ and φ components of Reynolds stresses should not lead to changes in the total angular momentum, the last term in (5.24) retains only the component describing friction against the Earth (under the assumption that the vertical turbulent flux of momentum at the upper atmosphere boundary is zero). On denoting this component by $\tau_\lambda |_{\sigma=1}$ we shall finally have

$$\frac{\partial}{\partial t} \int_0^1 \int_G \pi M \, dG \, d\sigma = \int_G \left(\Phi_S \frac{\partial \pi}{\partial \lambda} + a \cos \varphi \, \tau_\lambda |_{\sigma=1} \right) dG. \tag{5.25}$$

Averaging (5.25) over a sufficiently large time interval and assuming that changes in the pressure π are determined mainly by the changes in Φ_S (dynamic changes in π are significantly smaller) it

is possible to obtain the approximate relation

$$\int_G \overline{\tau_\lambda|_{\sigma=1}\, a\cos\varphi}^{t}\, dG \approx 0 \tag{5.26}$$

(the bar with the index t stands for time averaging). Indeed in this case we shall have

$$\int_G \overline{\tau_\lambda|_{\sigma=1}\, a\cos\varphi}^{t}\, \partial G = -\int_G (\partial\pi/\partial\lambda)\,\Phi_S\, dG = \int_G \pi\,(\partial\Phi_S/\partial\lambda)\, dG$$

$$= \int_G \pi\,(\Phi_S)\,(\partial\Phi_S/\partial\lambda)\, dG = \int_G [\partial\psi\,(\Phi_S)/\partial\lambda]\, dG = 0,$$

since $\psi\,(\Phi_S)$ is a periodic function of λ. Here

$$\pi\,(\Phi_S) = \partial\psi\,(\Phi_S)/\partial\Phi_S.$$

3. Similarly one can obtain the law of the conservation of specific humidity

$$\frac{\partial}{\partial t}\int_0^1\int_G \pi q\, dG\, d\sigma = \int_G \left[E_S - \int_0^1 \pi\,(C-E)\, d\sigma\right] dG, \tag{5.27}$$

where E_S is the humidity flux from the Earth's surface. Assuming that all the condensed moisture precipitates (that is ignoring the evaporation of drops in the clouds and taking $E = 0$) the right-hand side of (5.27) expresses the balance between the amount of moisture precipitated and the amount of moisture evaporated from the surface of the land and ocean.

4. We now obtain the law of the conservation of total energy (subject to the absence of dissipation and influxes). By multiplying (5.1) by πu and (5.2) by πv, adding the resulting expressions, and taking into account the continuity equation (5.4) we have the following equation for the kinetic energy $K = (u^2 + v^2)/2$:

$$\frac{\partial(\pi K)}{\partial t} + \frac{1}{a\cos\varphi}\left(\frac{\partial(u\pi K)}{\partial\lambda} + \frac{\partial(v\cos\varphi\,\pi K)}{\partial\varphi}\right) + \frac{\partial(\dot\sigma\pi K)}{\partial\sigma}$$

$$+\pi\left[\frac{1}{a\cos\varphi}\left(\frac{\partial\Phi}{\partial\lambda} + \frac{RT}{\pi}\frac{\partial\pi}{\partial\lambda}\right) + \frac{v}{a}\left(\frac{\partial\Phi}{\partial\varphi} + \frac{RT}{\pi}\frac{\partial\pi}{\partial\varphi}\right)\right]$$

$$= \pi\,(uF_u + vF_v). \tag{5.28}$$

In (5.28) we transform the terms involving the geopotential gradient:

$$\pi\left(\frac{u}{a\cos\varphi}\frac{\partial\Phi}{\partial\lambda} + \frac{v}{a}\frac{\partial\Phi}{\partial\varphi}\right) = \frac{1}{a\cos\varphi}\left[\frac{\partial(u\pi\Phi)}{\partial\lambda} + \frac{\partial(v\cos\varphi\,\pi\Phi)}{\partial\varphi}\right.$$

$$\left. -\Phi\left(\frac{\partial(\pi u)}{\partial\lambda} + \frac{\partial(\pi v\cos\varphi)}{\partial\varphi}\right)\right].$$

By virtue of (5.4)

$$-\frac{\Phi}{a\cos\varphi}\left(\frac{\partial(\pi u)}{\partial\lambda}+\frac{\partial(\pi v\cos\varphi)}{\partial\varphi}\right)=\Phi\left(\frac{\partial\pi}{\partial t}+\frac{\partial(\pi\dot{\sigma})}{\partial\sigma}\right). \tag{5.29}$$

The expression on the right-hand side of (5.29) can be written as

$$\Phi\left(\frac{\partial\pi}{\partial t}+\frac{\partial(\pi\dot{\sigma})}{\partial\sigma}\right)=\frac{\partial}{\partial\sigma}\left(\sigma\frac{\partial\pi}{\partial t}+\pi\dot{\sigma}\right)\Phi-\left(\sigma\frac{\partial\pi}{\partial t}+\pi\dot{\sigma}\right)\frac{\partial\Phi}{\partial\sigma}.$$

And, finally, on using equation (5.3) we get

$$-\left(\sigma\frac{\partial\pi}{\partial t}+\pi\dot{\sigma}\right)\frac{\partial\Phi}{\partial\sigma}+\left(\sigma\frac{\partial\pi}{\partial t}+\pi\dot{\sigma}\right)\frac{RT}{\sigma}.$$

As a result of the transformations equation (5.28) assumes the form

$$\frac{\partial\pi K}{\partial t}+\frac{1}{a\cos\varphi}\left(\frac{\partial(u\pi K)}{\partial\lambda}+\frac{\partial(v\cos\varphi\,\pi K)}{\partial\varphi}\right)+\frac{\partial(\dot{\sigma}\pi K)}{\partial\sigma}$$
$$+\frac{1}{a\cos\varphi}\left(\frac{\partial(u\pi\Phi)}{\partial\lambda}+\frac{\partial(v\cos\varphi\,\pi\Phi)}{\partial\varphi}\right)+\frac{\partial}{\partial\sigma}\left(\sigma\frac{\partial\pi}{\partial t}+\pi\dot{\sigma}\right)\Phi$$
$$+\frac{RT}{\sigma}\left[\pi\dot{\sigma}+\sigma\left(\frac{\partial\pi}{\partial t}+\frac{u}{a\cos\varphi}\frac{\partial\pi}{\partial\lambda}+\frac{v}{a}\frac{\partial\pi}{\partial\varphi}\right)\right]=\pi(uF_u+vF_v). \tag{5.30}$$

Now add equation (5.5) multiplied by π and (5.30) and integrate the resulting expression for σ going from 0 to 1 and with respect to the entire Earth's surface. Using the corresponding boundary conditions we shall have

$$\frac{\partial}{\partial t}\int_G \pi\left[\Phi_S+\int_0^1 (K+c_pT)\,d\sigma\right]dG$$
$$=\int_G\int_0^1 \pi(uF_u+vF_v+F_T+\varepsilon)\,d\sigma\,dG. \tag{5.31}$$

In the adiabatic approximation ($F_u=F_v=F_T=\varepsilon=0$) a law of the conservation of the system's total energy is thus valid.

Note one more integral law of conservation which can be obtained under a simplifying assumption. Specifically, in the heat influx equation we replace the term $RT/(c_p\sigma)$ by $RT^*/(c_p\sigma)$, where T^* is the temperature mean value, a quantity independent of space coordinates and time. It is then easy to see that in the adiabatic approximation the law of the conservation of total energy is valid in the form

$$\frac{\partial}{\partial t}\int_G \pi\left[\Phi_S+\int_0^1\left(K+\frac{c_p}{2T^*}T^2\right)d\sigma\right]dG=0. \tag{5.32}$$

For $\Phi_S = 0$ the energy is a quadratic form of the quantities $\sqrt{\pi}u$, $\sqrt{\pi}v$, $\sqrt{\pi}T$.

5.2.2. The Ocean. We now proceed to consider conservation laws for the equations of the hydrothermodynamics of the ocean. Assuming that there are no frictional force and external sources in the ocean and writing equations (5.11) to (5.17) in terms of real quantities ρ, T, and S we arrive at a system of equations similar to (5.11) to (5.16), with γ_T and γ_S zero. We shall assume that the lateral cylindrical surface Γ of the ocean basin Ω is a no-flow surface, i.e.

$$(\mathbf{u}, \mathbf{n}) = 0 \quad \text{on } \Gamma, \tag{5.33}$$

where $\mathbf{u}$ is the horizontal velocity vector with components u and v and $\mathbf{n}$ is the direction of the outward normal to Γ.

The heat- and salt-transfer equations yield the conservation laws

$$\frac{\partial}{\partial t}\int_{\Omega} HT\, d\Omega = 0, \quad \frac{\partial}{\partial t}\int_{\Omega} HS\, d\Omega = 0, \tag{5.34}$$

where $d\Omega$ is the elementary volume of a three-dimensional region Ω:

$$d\Omega = a^2 \cos\varphi\, d\lambda\, d\varphi\, dz.$$

We establish a basic law of the conservation of the total energy of system (5.11) to (5.14), (5.33) similar to the one that held for the atmosphere. We first reduce the two equations, (5.15) and (5.16), (where $\gamma_T = \gamma_S = 0$, $F_{ST} = F_{SS} = 0$), to the equation for the sea water density taking into account the equation of state [5.6]

$$\rho = \rho_0 - \alpha_0 T + \beta_0 S - \gamma_0 TS - \sigma_0 T^2,$$

$$\rho_0 = 1.000082, \quad \alpha_0 = 35 \times 10^{-7}, \quad \beta_0 = 802 \times 10^{-6},$$

$$\gamma_0 = 2 \times 10^{-6}, \quad \sigma_0 = 469 \times 10^{-8}.$$

On multiplying the equations for T and S respectively pairwise by $(-\alpha_0, \beta_0)$, $(-\gamma_0 S, -\gamma_0 T)$ and $(-\sigma_0 T, 0)$ and adding up the results we get

$$H\, d\rho/dt = 0. \tag{5.35}$$

Now multiply (5.11) by $\rho_0 u$, (5.12) by $\rho_0 v$, and (5.35) by $-gz$. Adding together all the equations and carrying out integration over Ω, we obtain the law of the conservation of the total energy of system (5.11)-(5.14), (5.35), (5.33):

$$\frac{\partial}{\partial t}\int_{\Omega} H\left(\rho_0 \frac{u^2+v^2}{2} - gz\rho\right) d\Omega = 0, \tag{5.36}$$

where the first term in the integrand represents the kinetic energy and the second term represents the potential energy.

For the original system (5.11)-(5.17) written in terms of ρ', T', S' (γ_T, γ_S in (5.15), (5.16) are nonzero), with the additional assumption that the density field is linearly dependent on the temperature and salinity, $\rho = \bar{\rho} - \alpha_T T' + \beta_S S'$, $\bar{\rho} = \hat{\rho}_0 - \alpha_T \bar{T}(z) + \beta_S \bar{S}(z)$, $\hat{\rho}_0 = \text{const}$, we have the following quadratic law of conservation:

$$\frac{\partial}{\partial t} \int_{\Omega} H \left(\hat{\rho}_0 \frac{u^2+v^2}{2} + \frac{g(\rho')^2}{2\gamma} \right) d\Omega = 0 \qquad (5.37)$$

where $\rho' = \rho - \bar{\rho}$, $\gamma = -\alpha_T \gamma_T + \beta_S \gamma_S > 0$, $\gamma_T = d\bar{T}/dz$, $\gamma_S = d\bar{S}/dz$.

Some words are in order concerning asymptotic laws of conservation which appear to affect significantly the statistical characteristics of the solution of the system of equations for the dynamics of the ocean and atmosphere. If one considers a barotropic incompressible fluid, this has an infinite number of invariants obeying local laws of the conservation of enstrophy. Two quadratic invariants (energy and enstrophy) prohibit the cascade of energy over the spectrum in the direction of high wave numbers, thus determining the slope of the gain-frequency characteristic for the kinetic energy of the system at the inertia scale of waves (the so-called law of power -3). From the point of view of difference schemes this prohibition of the energy cascade plays an important part where there are powerful sources of generating waves of various scale, such as orographic inhomogeneities of the Earth's surface. In the case of difference schemes with a single quadratic invariant (energy), if there is a generating source one has to use additional artificial "viscosity" to compensate for the excessive energy cascade over the spectrum in the direction of short waves. It is not completely understood what form the artificial viscosity must have to describe correctly the transformation of energy over the spectrum. It seems highly problematical at least to use operators of the Laplacian type.

Another aspect of the problem is the possibility for a system having one quadratic invariant of the type of energy integral to become stochastic. Although we have not a single quadratic invariant in the complete systems of equations for both the atmosphere and the ocean, it is natural to assume that if a set of quadratic laws of conservation holds for some truncated subsystems this will stabilize the original systems, especially in the presence of constantly acting forcings. The use of schemes with a single exact quadratic invariant allows, however, solving a number of special problems, for example, the problem of forming a space and a time spectrum of the solution and the relation between those spectra. The formation of a difference scheme that has the necessary properties is naturally connected with

special forms of the original systems of differential equations.*

Before proceeding to the construction of numerical schemes for the solution of the given problems in the dynamics of the ocean and atmosphere, we will carry out some transformations that reduce the original systems (5.1)-(5.6) and (5.11)-(5.17) to the most convenient form.

5.3. THE SYMMETRIZED FORM OF THE EQUATIONS OF THE ATMOSPHERE DYNAMICS AND THE EVOLUTIONARY FORMULATION OF THE PROBLEM IN THE OCEAN DYNAMICS

5.3.1. The Atmosphere. Construction of absolutely stable difference schemes for a model of the atmosphere dynamics (in the adiabatic approximation) seems to be most easily carried out for the so-called symmetrized form of the equations. Such a symmetrization of the equations of the atmosphere dynamics was reported for the first time in [5.7]. Note that the symmetrized form of the equations enables an infinite number of absolutely stable schemes to be constructed. Indeed, consider, for example, the two equations,

$$\frac{\partial T}{\partial t}+u\frac{\partial T}{\partial x}=0,\quad \frac{\partial \pi}{\partial t}+\frac{\partial (\pi u)}{\partial x}=0,\quad x\in\Omega,\quad t\in[0,\ T], \tag{5.38}$$

which are a simplified variant of the heat transfer equation and of the continuity equation. Both can be reduced to the form

$$\begin{gathered}\frac{\partial (\sqrt{\pi}\,T)}{\partial t}+\frac{1}{2}u\frac{\partial (\sqrt{\pi}\,T)}{\partial x}+\frac{1}{2}\frac{\partial (\sqrt{\pi}\,uT)}{\partial x}=0,\\ \frac{\partial \sqrt{\pi}}{\partial t}+\frac{1}{2}\frac{\partial (\sqrt{\pi}\,u)}{\partial x}+\frac{1}{2}u\frac{\partial \sqrt{\pi}}{\partial x}=0.\end{gathered} \tag{5.39}$$

Changing the variables $\psi=\sqrt{\pi}T$, $\varphi=\sqrt{\pi}$ we get

$$\begin{gathered}\partial\psi/\partial t+(1/2)\,u\,\partial\psi/\partial x+(1/2)\,\partial(\psi u)/\partial x=0,\\ \partial\varphi/\partial t+(1/2)\,u\,\partial\varphi/\partial x+(1/2)\,\partial(\varphi u)/\partial x=0.\end{gathered} \tag{5.40}$$

These equations are absolutely equivalent in point of form, so it is possible to consider just one of them. It is easy to see that for (5.40) we have

$$\frac{\partial}{\partial t}\int\limits_{\Omega}\psi^2\,d\Omega=0$$

*For example, when constructing difference schemes having in the shallow water approximation energy and enstrophy conservation laws, it is convenient to use the system of equations in the so-called Lamb form.

provided u vanishes on the boundary of Ω, or ψ is periodic. This follows from the skew-symmetry of the operator

$$K\varphi \equiv u\,\partial\varphi/\partial x + \partial\,(u\varphi)/\partial x.$$

It is easy to verify that the finite-difference approximations of the operator K will also preserve the skew-symmetry condition (a skew-symmetric matrix will form) if for each term $u\,\partial\psi/\partial x$ and $\partial\,(\psi u)/\partial x$ in (5.40) we use the same uniform symmetric pattern. Then we shall have

$$\partial\psi^h/\partial t + K_h\psi^h = 0 \quad \text{provided} \quad (K_h\psi^h,\ \psi^h) = 0,$$

where the scalar product $(\varphi^h,\ \psi^h) = \sum_i \varphi_i^h\psi_i^h$ is chosen in the corresponding grid space.

We make an important remark. The skew-symmetric form of the operator K_h will be preserved regardless of what we choose as u^h, i.e. it is independent of the form of the projector that carries u into u^h. This independence of the skew-symmetry property of the form of the projection operator P_h enables a family of schemes for (5.40) that satisfy $(K_h\psi^h,\ \psi^h) = 0$ to be at once constructed. If we now use the Crank-Nicholson scheme with respect to time, we shall have

$$[(\psi^h)^{j+1} - (\psi^h)^j]/\tau + K_h\,[(\psi^h)^{j+1} + (\psi^h)^j]/2 = 0. \tag{5.41}$$

Performing the scalar multiplication of (5.41) by $[(\psi^h)^{j+1} + (\psi^h)^j]/2$ we get

$$((\psi^h)^{j+1},\ (\psi^h)^{j+1})) = ((\psi^h)^j,\ (\psi^h)^j), \tag{5.42}$$

which means that problem (5.41) is stable.

At this point the question arises: which scheme of the family of difference schemes is to be preferred? First note that if we set $T = \text{const}$ in (5.39), the equation for ψ goes over into the one for φ. Since the equations are invariant under the translation group, we must generally require that the above condition hold for the difference scheme as well. Further, equations (5.39), besides preserving the quadratic form $\int_\Omega \pi T^2\,d\Omega$, have the linear laws $\int_\Omega \pi T\,d\Omega = \text{const}$ which are also physically meaningful (see the section on conservation laws). Therefore we must require that the first equation of (5.40) could be reduced with the aid of the second equation of (5.40) to the form that provides a linear conservation law.

If we change the variables

$$\sqrt{\pi} = p^*,\quad \sqrt{\pi}\,u = U,\quad \sqrt{\pi}\,v = V,\quad \sqrt{\pi}\,T = \Theta,\ \sqrt{\pi}q = Q, \tag{5.43}$$

the symmetrized form of the equations of hydrothermodynamics will become

$$\frac{\partial U}{\partial t}+KU-\left(1+\frac{U}{ap^*}\tan\varphi\right)V^{\cdot}$$
$$+\frac{1}{a\cos\varphi}\left[\frac{p^*\partial\Phi}{\partial\lambda}+\frac{2R\Theta}{p^*}\frac{\partial p^*}{\partial\lambda}\right]=\frac{F_u}{p^*},$$
$$\frac{\partial V}{\partial t}+KV+\left(1+\frac{U}{ap^*}\tan\varphi\right)U$$
$$+\frac{1}{a}\left(p^*\frac{\partial\Phi}{\partial\varphi}+\frac{2R\Theta}{p^*}\frac{\partial p^*}{\partial\varphi}\right)=\frac{F_v}{p^*},$$
$$\frac{\partial\Theta}{\partial t}+K\Theta-\frac{2R\Theta}{c_p}\left(\frac{\partial p^*}{\partial t}+\frac{U}{p^*a\cos\varphi}\frac{\partial p^*}{\partial\lambda}+\frac{V}{p^*a}\frac{\partial p^*}{\partial\varphi}\right)$$
$$-\frac{R\Theta p^*}{c_p\sigma}\dot{\sigma}=\frac{F_T}{p^*}+\frac{\varepsilon}{p^*},$$
$$\frac{\partial\Phi}{\partial\sigma}=-\frac{R\Theta}{p^*\sigma},\qquad\frac{\partial p^*}{\partial t}+Kp^*=0, \tag{5.44}$$
$$\frac{\partial Q}{\partial t}+KQ=\frac{F_q}{p^*}-\frac{C-E}{p^*},$$
$$K=\frac{1}{2a\cos\varphi}\left[\frac{U}{p^*}\frac{\partial}{\partial\lambda}+\frac{\partial(U/p^*)}{\partial\lambda}+\frac{V\cos\varphi}{p^*}\frac{\partial}{\partial\varphi}+\frac{\partial}{\partial\varphi}\frac{V\cos\varphi}{p^*}\right]$$
$$+\frac{1}{2}\dot{\sigma}\frac{\partial}{\partial\sigma}+\frac{1}{2}\frac{\partial\dot{\sigma}}{\partial\sigma}.$$

Note that a different form of symmetrization can be introduced for the specific humidity q by setting $\sqrt{\pi q}=Q$. (This is possible since π and q are nonnegative.) Then the equation assumes the form

$$\frac{\partial Q}{\partial t}+KQ=\frac{F_q}{2Q}-\frac{C-E}{2Q}$$

and the linear conservation law for q goes automatically over into the quadratic one for Q.

5.3.2. The Ocean. If one compares the two statements of the problem of the large-scale circulation of the atmosphere, (5.1) to (5.6), and the ocean, (5.11) to (5.17), one can note that mathematically the main differences between them that determine the peculiarity of the choice of the form of equations and algorithms for their numerical solution are as follows. First, as already pointed out, the statement of the problem of the atmosphere dynamics has a specific feature (the number of equations in it is a unit less than that of the desired functions) and the continuity equation (5.4) is of evolutionary form. Second, the problem of the ocean dynamics is a boundary-value problem in variables λ and φ.

We have chosen the method of splitting up by physical processes and geometrical variables [5.8], [5.9] as the basic method for inte-

grating with respect to time the equations of the dynamics of the ocean and atmosphere. To justify the use of the splitting method it is necessary that the system of equations be written in evolutionary form and that the operator of the problem be represented as the sum of positive or positive definite operators of a simpler form. Of importance in choosing the form of the equations of the atmosphere dynamics is, therefore, the above symmetrization of the original system, after which the operator of the problem is rather simply broken down into sums of positive operators and it is simple to reduce the problem to the evolutionary form*.

In the problem of the dynamics of the ocean the transformation of the original system of equations (5.11) to (5.17) to the evolutionary form comes to the fore and the question of how to split the operator of the problem is easier to tackle.

To reduce problem (5.11)-(5.17) to the evolutionary form we proceed as follows. Represent the horizontal components of the velocity vector as

$$u = \tilde{u} + u', \quad v = \tilde{v} + v', \tag{5.45}$$

where $\sim$ signifies averaging over the vertical: $\tilde{a} = \int_0^1 a\, dz$. Substitute (5.45) into (5.11), (5.12) and integrate them from the surface $z_1 = 0$ to the bottom $z_1 = 1$. Using the hydrostatics equation

$$p(z) = p_0 + gH \int_0^z \rho\, dz$$

we get

$$H\frac{\partial \tilde{u}}{\partial t} + \frac{H\tilde{u}}{a\cos\varphi}\frac{\partial \tilde{u}}{\partial \lambda} + \frac{H\tilde{v}}{a}\frac{\partial \tilde{u}}{\partial \varphi} - H\left(l + \frac{\tan\varphi}{a}\tilde{u}\right)\tilde{v}$$

$$= -\frac{H}{a\cos\varphi\,\rho_0}\frac{\partial p_0}{\partial \lambda} - \frac{1}{a\cos\varphi\,\rho_0}\int_0^1 \left(Hg\frac{\partial}{\partial \lambda}H\int_0^z \rho\, dz - \frac{\partial H}{\partial \lambda}zgH\rho\right)dz$$

$$+ \int_0^1 HF_{Su}\, dz + J_\lambda, \tag{5.46}$$

*Taking into account the additional boundary condition (5.10), Φ is found from (5.3) and the equation of hydrostatics, which is not an evolutionary one, is thus eliminated from the system. Equation (5.4) together with conditions (5.8) and (5.9) is integrated over the vertical coordinate, giving rise to the equation for π containing no dependence from the vertical velocity $\dot{\sigma}$. Finally, from the remaining equations (5.1), (5.2), (5.5), and (5.6) $\dot{\sigma}$ is eliminated using integration of (5.4), with the constructed equation for π taken into account.

$$H\frac{\partial\tilde{v}}{\partial t}+\frac{H\tilde{u}}{a\cos\varphi}\frac{\partial\tilde{v}}{\partial\lambda}+\frac{H\tilde{v}}{a}\frac{\partial\tilde{v}}{\partial\varphi}+H\left(l+\frac{\tan\varphi}{a}\tilde{u}\right)\tilde{u} \tag{5.46}$$

$$=-\frac{H}{a\rho_0}\frac{\partial p_0}{\partial\varphi}-\frac{1}{a\rho_0}\int_0^1\left(Hg\frac{\partial}{\partial\varphi}H\int_0^z\rho\,dz-\frac{\partial H}{\partial\varphi}zgH\rho\right)dz$$

$$+\int_0^1 HF_{Sv}\,dz+J_\varphi.$$

In addition, from the continuity equation we have

$$\frac{\partial(H\tilde{u})}{\partial\lambda}+\frac{\partial}{\partial\varphi}(H\tilde{v}\cos\varphi)=0. \tag{5.47}$$

The terms J_λ and J_θ describe the effect of the baroclinic components of the horizontal flux on the plane circulation:

$$J_\lambda=\int_0^1\left[\sin\varphi u'v'-\frac{1}{a\cos\varphi}\frac{\partial}{\partial\lambda}(H(u')^2)-\frac{1}{a}\frac{\partial}{\partial\varphi}(Hu'v')\right]dz, \tag{5.48}$$

$$J_\varphi=\int_0^1\left[-\sin\varphi(u')^2-\frac{1}{a\cos\varphi}\frac{\partial}{\partial\lambda}(Hu'v')-\frac{1}{a}\frac{\partial}{\partial\varphi}(H(v')^2)\right]dz.$$

Subtracting equations (5.46) and (5.47) from (5.11) and (5.12) and from the continuity equation we get

$$H\frac{\partial u'}{\partial t}+\frac{H}{a\cos\varphi}(\tilde{u}+u')\frac{\partial u'}{\partial\lambda}+\frac{H}{a}(\tilde{v}+v')\frac{\partial u'}{\partial\varphi}+w\frac{\partial u'}{\partial z}$$

$$-H\left[l+(\tilde{u}+u')\frac{\tan\varphi}{a}\right]v'=-\frac{1}{a\cos\varphi\rho_0}\left(I_\lambda-\int_0^1 I_\lambda\,dz\right)$$

$$+HF_{Su}-\int_0^1 HF_{Su}\,dz-(N_\lambda+J_\lambda),$$

$$H\frac{\partial v'}{\partial t}+\frac{H}{a\cos\varphi}(\tilde{u}+u')\frac{\partial v'}{\partial\lambda}+\frac{H}{a}(\tilde{v}+v')\frac{\partial v'}{\partial\varphi}+w\frac{\partial v'}{\partial z} \tag{5.49}$$

$$+H\left[l+(\tilde{u}+u')\frac{\tan\varphi}{a}\right]u'=-\frac{1}{a\rho_0}\left(I_\varphi-\int_0^1 I_\varphi\,dz\right)$$

$$+HF_{Sv}-\int_0^1 HF_{Sv}\,dz-(N_\varphi+J_\varphi),$$

$$H\frac{\partial T}{\partial t}+\frac{H}{a\cos\varphi}(\tilde{u}+u')\frac{\partial T}{\partial\lambda}+\frac{H}{a}(\tilde{v}+v')\frac{\partial T}{\partial\varphi}+w\frac{\partial T}{\partial z}$$
$$=\frac{1}{H}\frac{\partial}{\partial z}\nu_T\frac{\partial T}{\partial z}+HF_{ST}-\gamma_T w_2,$$

$$H\frac{\partial S}{\partial t}+\frac{H}{a\cos\varphi}(\tilde{u}+u')\frac{\partial S}{\partial\lambda}+\frac{H}{a}(\tilde{v}+v')\frac{\partial S}{\partial\varphi}+w\frac{\partial S}{\partial z}$$
$$=\frac{1}{H}\frac{\partial}{\partial z}\nu_S\frac{\partial S}{\partial z}+HF_{SS}-\gamma_S w_2,$$

$$w=\int_z^1\frac{1}{a\cos\varphi}\left[\frac{\partial(Hu')}{\partial\lambda}+\frac{\partial}{\partial\varphi}(Hv'\cos\varphi)\right]dz, \tag{5.49}$$

$$I_\lambda=H\frac{\partial}{\partial\lambda}\left(gH\int_0^z\rho\,dz\right)-\frac{\partial H}{\partial\lambda}zgH\rho,$$

$$I_\varphi=H\frac{\partial}{\partial\varphi}\left(gH\int_0^z\rho\,dz\right)-\frac{\partial H}{\partial\varphi}zgH\rho,$$

$$N_\lambda=\frac{H}{a\cos\varphi}\frac{\partial\tilde{u}}{\partial\lambda}u'+\frac{H}{a}\frac{\partial\tilde{u}}{\partial\varphi}v'-H\frac{\tan\varphi}{a}u'\tilde{v},$$

$$N_\varphi=\frac{H}{a\cos\varphi}\frac{\partial\tilde{v}}{\partial\lambda}u'+\frac{H}{a}\frac{\partial\tilde{v}}{\partial\varphi}v'+H\frac{\tan\varphi}{a}u'\tilde{u}.$$

Finally, the last step in the transformations is to eliminate the function p_0 from (5.46). To do this we use (5.47) and introduce the integral stream function

$$\frac{1}{H}\frac{\partial\psi}{\partial\varphi}=\tilde{u},\qquad\frac{1}{H\cos\varphi}\frac{\partial\psi}{\partial\lambda}=-\tilde{v}. \tag{5.50}$$

Next we multiply the first equation of (5.46) by $\cos\varphi/H$ and the second by $1/H$ and perform the vorticity operation for them. We get

$$H\frac{\partial}{\partial t}\left(\frac{A\psi}{H}\right)+\frac{H}{a\cos\varphi}\tilde{u}\frac{\partial}{\partial\lambda}\left(\frac{A\psi}{H}-\frac{l}{H}\right)+\frac{H}{a}\tilde{v}\frac{\partial}{\partial\varphi}\left(\frac{A\psi}{H}-\frac{l}{H}\right)$$
$$=\frac{1}{a\cos\varphi}\left(\frac{\partial}{\partial\varphi}\frac{\cos\varphi}{H}\int_0^1 HF_{Su}\,dz-\frac{\partial}{\partial\lambda}\frac{1}{H}\int_0^1 HF_{Sv}\,dz+Q\right), \tag{5.51}$$

$$Q=-\frac{g}{a\rho_0}\left[\frac{\partial}{\partial\varphi}\int_0^1\left(\frac{\partial}{\partial\lambda}H\int_0^z\rho\,dz-\frac{\partial H}{\partial\lambda}z\rho\right)dz\right.$$
$$\left.-\frac{\partial}{\partial\lambda}\int_0^1\left(\frac{\partial}{\partial\varphi}H\int_0^z\rho\,dz-\frac{\partial H}{\partial\varphi}z\rho\right)dz\right]+\frac{\partial}{\partial\varphi}\frac{\cos\varphi}{H\rho_0}(\tau_1+J_\lambda)$$
$$-\frac{\partial}{\partial\lambda}\frac{1}{H\rho_0}(\tau_2+J_\varphi),$$

Equation (5.51) is written for the integral stream function in operator form, where

$$\frac{A\psi}{H}=\frac{1}{Ha\cos\varphi}\left(\frac{\partial}{\partial\varphi}\frac{\cos\varphi}{H}\frac{\partial\psi}{\partial\varphi}+\frac{\partial}{\partial\lambda}\frac{1}{H\cos\varphi}\frac{\partial\psi}{\partial\lambda}\right). \tag{5.52}$$

As a result of the transformations the original equations (5.11) to (5.17) have been reduced to the evolutionary system (5.49), (5.51) for the functions u', v', T, S and ψ.

5.4. THE CHOICE OF GRIDS. APPROXIMATION OF EQUATIONS IN SPACE VARIABLES

In constructing finite-difference approximations of the systems of equations of the dynamics of the ocean and atmosphere, of great importance is the choice of domains of the desired grid functions. This is related to the fact that actual numerical calculations are carried out with finite (moreover, sufficiently large) spacings of a difference grid and the above properties of a solution of the original differential problem must hold to a good accuracy. This mainly concerns the optimal reproduction of the major physical characteristics of the model. It should include in the first place the following:

a description of the phase and group velocities of inertial gravity waves (especially long waves),

a reproduction of the stability criteria of baroclinic waves and the rates of growth of unstable modes,

a description as precise as possible of orographic effects,

a reproduction of the "locally skew-symmetric" character of Coriolis forces,

the most convenient representation of the dynamic operator at the pole points,

a convenient implementation of the boundary conditions.

In [5.10], [5.11] the various aspects of the choice of grid spaces have been studied from the point of view of satisfying the separate requirements formulated above. On the basis of the results obtained there it can be concluded that the most acceptable for both the model of the atmosphere and that of the ocean is a displaced grid in which the horizontal components of the velocity vector are in the middles of the squares at integer levels in the vertical coordinate and the other desired functions are at the grid points. The analogues of the vertical velocity are displaced by half a spacing in height with respect to the other functions. The horizontal boundaries of the grid region (for the ocean model) and the vertical boundaries (for both models) are displaced by half a spacing away from the basic grid; kinematic conditions are thus exactly satisfied at the boundaries of the grid region.

We now proceed to describe the methods of approximating the systems of equations of the dynamics of the ocean and atmosphere in space variables. Particular attention will be given to the fundamental points in the construction of numerical schemes. The reader interested in the details is referred to [5.1]-[5.4], [5.12], [5.13].

5.4.1. The Atmosphere. As the method of solving the problems of the dynamics of the ocean and atmosphere with respect to time we use the splitting method by physical processes and geometrical variables [5.8], [5.9]. We first split the discrete analogues of the transformed systems (5.44) and (5.49), (5.51) into two main stages describing the processes of the transport of substances and of the adaptation of hydrodynamic fields. The operators of the problem are then split within each stage according to individual coordinates. It is convenient therefore to develop the basic devices of spatial approximation separately for the system of equations describing the transport of substances along the trajectories and that describing the adaptation of the fields.

For the transport equation the construction of a difference scheme that satisfies the requirements formulated at the beginning of this section will be carried out using an example of a one-dimensional problem. We have

$$\frac{\partial \varphi}{\partial t} + \frac{1}{2} u \frac{\partial \varphi}{\partial x} + \frac{1}{2} \frac{\partial (\varphi u)}{\partial x} = 0, \tag{5.53}$$

where $\varphi = \sqrt{\pi} T$, and for $\sqrt{\pi}$ we have

$$\frac{\partial \sqrt{\pi}}{\partial t} + \frac{1}{2} u \frac{\partial \sqrt{\pi}}{\partial x} + \frac{1}{2} \frac{\partial (\sqrt{\pi} u)}{\partial x} = 0.$$

Consider the displaced grids on which u and $\sqrt{\pi} T$ are defined (Fig. 5.1). We write a scheme of the second order of accuracy with

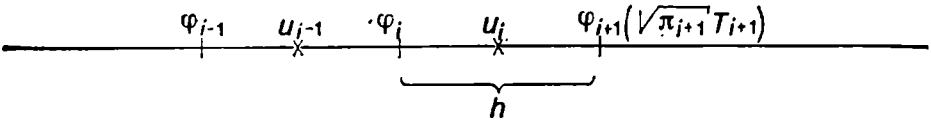


Fig. 5.1. Difference grid

respect to space for equation (5.53)

$$\frac{\partial \varphi_i}{\partial t} + \frac{1}{2} \tilde{u}_i \frac{\varphi_{i+1} - \varphi_{i-1}}{2h} + \frac{1}{2} \frac{\varphi_{i+1} \tilde{u}_{i+1} - \varphi_{i-1} \tilde{u}_{i-1}}{2h} = 0,$$

where $\tilde{u}_i$ are determined at the same points as φ_i are. We write this equation in the form

$$\frac{\partial \varphi_i}{\partial t} + \frac{[(\tilde{u}_{i+1} + \tilde{u}_i)/2]\, \varphi_{i+1} - [(\tilde{u}_i + \tilde{u}_{i-1})/2]\, \varphi_{i-1}}{2h} = 0.$$

Since we have displaced grids, it may be assumed, with the order of accuracy preserved, that

$$(\tilde{u}_{i+1} + \tilde{u}_i)/2 = u_i, \quad (\tilde{u}_i + \tilde{u}_{i-1})/2 = u_{i-1}.$$

Then we get

$$\partial\varphi_i/\partial t + (u_i\varphi_{i+1} - u_{i-1}\varphi_{i-1})/2h = 0. \tag{5.54}$$

It is obvious that the spatial difference operator is skew-symmetric and we have

$$\frac{\partial}{\partial t}\sum \varphi_i^2 = 0$$

(provided the boundary conditions are periodic or $u = 0$ at the boundaries).

For scheme (5.54) to be invariant under the translation group (with respect to T) it is necessary that the condition

$$\frac{\partial \sqrt{\pi_i}\, c}{\partial t} + \frac{u_i \sqrt{\pi_{i+1}}\, c - u_{i-1}\sqrt{\pi_{i-1}}\, c}{2h} = 0, \quad c = \text{const},$$

hold, i.e. we automatically obtain a difference scheme for $\sqrt{\pi_i}$:

$$\frac{\partial \sqrt{\pi_i}}{\partial t} + \frac{u_i \sqrt{\pi_{i+1}} - u_{i-1}\sqrt{\pi_{i-1}}}{2h} = 0. \tag{5.55}$$

We show that in this case a linear conservation law will also hold. On multiplying (5.54) by $\sqrt{\pi_i}$ and (5.55) by φ_i and adding up we get

$$\sqrt{\pi_i}\,\frac{\partial\varphi_i}{\partial t} + \varphi_i \frac{\partial \sqrt{\pi_i}}{\partial t} + \sqrt{\pi_i}\,\frac{u_i\varphi_{i+1} - u_{i-1}\varphi_{i-1}}{2h} + \varphi_i \frac{u_i \sqrt{\pi_{i+1}} - u_{i-1}\sqrt{\pi_{i-1}}}{2h} = 0$$

or

$$\frac{\partial \sqrt{\pi_i}\,\varphi_i}{\partial t} - \frac{1}{2h}(\sqrt{\pi_i}\,u_i\varphi_{i+1} - \sqrt{\pi_i}\,u_{i-1}\varphi_{i-1}) + \frac{1}{2h}(u_i\varphi_i\sqrt{\pi_{i+1}} - u_{i-1}\varphi_i\sqrt{\pi_{i-1}}) = 0.$$

It is easy to notice that a shift in the subscript turns the first term in the parentheses into the fourth and the second into the third. Therefore when summed the expressions in the parentheses vanish and hence

$$\frac{\partial}{\partial t}\sum_i \sqrt{\pi_i}\,\varphi_i = 0,$$

that is,

$$\frac{\partial}{\partial t}\sum_i \pi_i T_i = 0$$

(bearing in mind the above boundary conditions).

We now proceed to the approximation in the time variable. We write the Crank-Nicholson scheme for equation (5.53):

$$\frac{\varphi_i^{j+1}-\varphi_i^j}{\tau}+\frac{u_i\varphi_{i+1}^{j+1/2}-u_{i-1}\varphi_{i-1}^{j+1/2}}{2h}=0, \tag{5.56}$$

where $\varphi^{j+1/2}=(\varphi^{j+1}+\varphi^j)/2$. Performing a scalar multiplication of (5.56) by $\varphi^{j+1/2}$ we obtain the relation

$$\sum_i (\varphi_i^{j+1})^2=\sum (\varphi_i^j)^2.$$

The invariance under the translation group is preserved if for the continuity equation ($\sqrt{\pi}$) we also use the Crank-Nicholson scheme.

It is not hard to show that in this case the linear conservation law will also hold.

Choosing as an example the transport equation for the temperature $\Theta=\sqrt{\pi}T$ we write its complete approximation in space variables:

$$\begin{aligned}\frac{\partial\Theta_{i,j,k}}{\partial t}&+\frac{1}{2a\cos\varphi_j\,\Delta\lambda}\left[\bar{u}_{i+1/2,j,k}\Theta_{i+1,j,k}-\bar{u}_{i-1/2,j,k}\Theta_{i-1,j,k}\right]\\&+\frac{1}{2a\cos\varphi_j\,\Delta\lambda}\left[(\bar{v}\cos\varphi)_{i,j+1/2,k}\Theta_{i,j+1,k}\right.\\&\qquad\left.-(\bar{v}\cos\varphi)_{i,j-1/2,k}\Theta_{i,j-1,k}\right]\\&+\frac{1}{2\Delta\sigma_k}\left[\dot{\sigma}_{i,j,k+1/2}\Theta_{i,j,k+1}-\dot{\sigma}_{i,j,k-1/2}\Theta_{i,j,k-1}\right]=0,\end{aligned} \tag{5.57}$$

$$\bar{u}_{i+1/2,j,k}=\frac{u_{i+1/2,j+1/2,k}+u_{i+1/2,j-1/2,k}}{2},$$

$$(\bar{v}\cos\varphi)_{i,j+1/2,k}=\frac{(v\cos\varphi)_{i+1/2,j+1/2,k}+(v\cos\varphi)_{i-1/2,j+1/2,k}}{2}.$$

We now proceed to construct a finite-difference approximation of the system of equations of the adaptation of meteorological fields. The evolutionary form of the adaptation stage is as follows:

$$\begin{aligned}&\frac{\partial u}{\partial t}-\left(l+\frac{u}{ap^*}\tan\varphi\right)v+\frac{1}{a\cos\varphi}\left[p^*\frac{\partial\Phi}{\partial\lambda}+\frac{2R\Theta}{p^*}\frac{\partial p^*}{\partial\lambda}\right]=0,\\&\frac{\partial v}{\partial t}+\left(l+\frac{u}{ap^*}\tan\varphi\right)u+\frac{1}{a}\left[p^*\frac{\partial\Phi}{\partial\varphi}+\frac{2R\Theta}{p^*}\frac{\partial p^*}{\partial\varphi}\right]=0,\\&\frac{\partial\Theta}{\partial t}-\frac{2R\Theta}{c_pap^*}\left(\frac{u}{\cos\varphi}\frac{\partial\ln p^*}{\partial\lambda}+v\frac{\partial\ln p^*}{\partial\varphi}\right)+\frac{2}{c_pp^*}\frac{R\Theta}{p^*\sigma}\int_0^{\sigma}D\,d\sigma=0,\\&\frac{\partial p^*}{\partial t}+\int_0^1 D\,d\sigma=0,\end{aligned} \tag{5.58}$$

where

$$D=\frac{1}{2a\cos\varphi}\left(\frac{u}{p^*}\frac{\partial p^*}{\partial\lambda}+\frac{\partial u}{\partial\lambda}+\frac{v\cos\varphi}{p^*}\frac{\partial p^*}{\partial\varphi}+\frac{\partial(v\cos\varphi)}{\partial\varphi}\right).$$

Now we write the spatial finite-difference analogues of the equations:

$$\frac{\partial u_{i+1/2,\,j+1/2,\,k}}{\partial t}-\left(l_{j+1/2}+\frac{u_{i+1/2,\,j+1/2,\,k}}{ap^*_{i+1/1,\,j+1/2}}\tan\varphi_{j+1/2}\right)v_{i+1/2,\,j+1/2,\,k}$$
$$+\frac{1}{(a\cos\varphi)_{j+1/2}\,\Delta\lambda}\left[p^*_{i+1/2,\,j+1/2}\nabla_i\Phi_{j+1/2,\,k}\right.$$
$$\left.+\frac{2R\Theta_{i+1/2,\,j+1/2,\,k}}{p^*_{i+1/2,\,j+1/2}}\nabla_i p^*_{j+1/2}\right]=0,$$

$$\frac{\partial v_{i+1/2,\,j+1/2,\,k}}{\partial t}+\left(l_{j+1/2}+\frac{u_{i+1/2,\,j+1/2,\,k}}{ap^*_{i+1/2,\,j+1/2}}\tan\varphi_{j+1/2}\right)u_{i+1/2,\,j+1/2,\,k}$$
$$+\frac{1}{a\Delta\varphi}\left[p^*_{i+1/2,\,j+1/2}\nabla_j\Phi_{i+1/2,\,k}+\frac{2R\Theta_{i+1/2,\,j+1/2,\,k}}{p^*_{i+1/2,\,j+1/2}}\nabla_j p^*_{i+1/2}\right]=0,$$

$$\frac{\partial\Theta_{i,\,j,\,k}}{\partial t}-\frac{R}{c_p a\cos\varphi_j p^*_{i,\,j}}\sum_{\nu=-1,\,0}\left[\Theta_{i+1/2+\nu,\,j,\,k}u_{i+1/2+\nu,\,j,\,k}\nabla_{i+\nu}(\ln p^*)_j\right.$$
$$\left.+\cos\varphi_{j+1/2+\nu}\Theta_{i,\,j+1/2+\nu,\,k}v_{i,\,j+1/2+\nu,\,k}\nabla_{j+\nu}(\ln p^*)_i\right] \tag{5.59}$$
$$-\frac{2}{c_p p^*_{i,\,j}\,\Delta\sigma_k}\sum_{\nu=-1,\,0}\left[(\Phi_{i,\,j,\,k+1/2+\nu/2}-\Phi_{i,\,j,\,k+\nu/2})\sum_{\mu=1}^{k+\nu}D_{i,\,j,\,\mu}\,\Delta\sigma_\mu\right]=0,$$

$$\frac{\partial p^*_{i,\,j}}{\partial t}+\sum_{k=1}^{k}D_{i,\,j,\,k}\,\Delta\sigma_k=0,$$

$$\Phi_{i,\,j,\,k}-\frac{1}{\Delta\sigma_k}\nabla_{k-1/2}(\Phi_{i,\,j}\sigma)=\frac{R\Theta_{i,\,j,\,k}}{p^*_{i,\,j}},$$

$$D_{i,\,j,\,k}=\frac{1}{2a\cos\varphi_j p^*_{i,\,j}}$$
$$\times\left\{\frac{1}{2}\sum_{\nu=-1,\,0}\left[\frac{u_{i+1/2+\nu,\,j,\,k}\nabla_{i+\nu}p^*_j}{\Delta\lambda}+\frac{(v\cos'\varphi)_{i,\,j+1/2+\nu,\,k}\nabla_{j+\nu}p^*_i}{\Delta\varphi}\right]\right.$$
$$\left.+p^*_{i,\,j}\left[\frac{1}{\Delta\lambda}\nabla_{i-1/2}u_{j,\,k}+\frac{1}{\Delta\varphi}\nabla_{j-1/2}(v\cos\varphi)_{i,\,k}\right]\right\}.$$

In (5.59) we have used the following notation:

$$\nabla_r(\ )=(\ )_{r+1}-(\ )_r,$$
$$\nabla_{r+1/2}(\ )=[(\ )_{r+1}+(\ )_r]/2,$$

where r is any of the subscripts, i, j, or k.

Recall that the north and the south pole are the points of defining scalars. The value of p^* at a pole results from the meridional mass flow from all the points of the circle of the latitude where the velocity components are defined. To simplify computations the pole may be represented as a group of points with subscript i. For each i the continuity equation is used. Once $\partial p^*/\partial t$ for each i has been calculated, the value of $\partial p^*/\partial t$ for the poles is calculated as arithmetical mean over all i. A similar procedure is applied to the heat influx equation. Let $\Phi_S = 0$ and $\Theta = \Theta^* = \text{const}$. Then, multiplying the first equation of (5.59) by $u_{i+1/2,j+1/2,k}$, the second by $v_{i+1/2,j+1/2,k}$, and the third by $\Theta_{i,j,k}$ and summing over all i, j, k we get

$$\frac{\partial}{\partial t}\frac{1}{2}\sum_{i,j,k}\left(u^2_{i+1/2,\,j+1/2,\,k}+v^2_{i+1/2,\,j+1/2,\,k}+\frac{c_p\Theta^2_{i,j,k}}{\Theta^*}\right)=0. \qquad (5.60)$$

Equation (5.60) implies that if we represent (5.59) in vector form

$$\frac{\partial\varphi}{\partial t}+K_A\varphi=0, \quad \text{where} \quad \varphi=\begin{pmatrix} u_{i+1/2,\,j+1/2,\,k} \\ v_{i+1/2,\,j+1/2,\,k} \\ \sqrt{c_p/\Theta^*}\Theta_{i,\,j,\,k}\end{pmatrix},$$

and introduce a scalar product in a natural way, then the operator K_A in that scalar product will be skew-symmetric, i.e. $(K_A, \varphi, \varphi) = 0$. Thus the above difference (according to space) approximations for the advective part of the symmetrized form of the system of equation of hydrodynamics and for the part that describes the adaptation of the fields of meteorological elements allows us to write the resulting system of differential-difference equations for a complete system of equations of hydrodynamics (in the adiabatic approximation) as follows:

$$\partial\varphi/\partial t + K\varphi = 0, \quad \text{where} \quad (K\varphi, \varphi) = 0. \qquad (5.61)$$

5.4.2. The Ocean. Similarly to the problem of the dynamics of the atmosphere the construction of numerical approximations describing the general circulation of the ocean will be carried out for the two main stages of splitting the problem by physical processes. Assuming that the operators of the horizontal eddy diffusion in the ocean are of the form of Laplace operators, the initial boundary-value problem of the diffusion-transport of hydrological substance in the λ-coordinate takes on the form

$$Ha\cos\varphi\frac{\partial\psi}{\partial t}+Hu\frac{\partial\psi}{\partial\lambda}=\frac{1}{a\cos\varphi}\frac{\partial}{\partial\lambda}\left(\mu H\frac{\partial\psi}{\partial\lambda}\right), \qquad (5.62)$$

$$\partial\psi/\partial n = 0 \quad \text{when} \quad \lambda = 0,\ L, \quad \psi = \psi^0 \quad \text{when} \quad t = 0.$$

This equation is specific in that it is one with a small parameter of the higher derivative. Since (5.62) is a boundary-value problem, this imposes additional requirements on the choice of a numerical scheme. It is desirable that the scheme should solve the problem with a sufficiently small coefficient μ. With this in mind it seems appropriate to employ the projection method. We state the problem in variational form. On representing an approximate solution of (5.62) in the form

$$\psi_N = \sum_{i=1}^{N} \psi_i \varkappa_i (\lambda)$$

we have

$$\int_0^L \left\{ \left[\frac{\partial}{\partial t} \left(Ha \cos \varphi \sum_{i=1}^{N} \psi_i \varkappa_i \right) + uH \frac{\partial}{\partial \lambda} \left(\sum_{i=1}^{N} \psi_i \varkappa_i \right) \right] \omega_j \right.$$

$$\left. + \frac{1}{a \cos \varphi} \mu H \frac{\partial}{\partial \lambda} \left(\sum_{i=1}^{N} \psi_i \varkappa_i \right) \frac{\partial \omega_j}{\partial \lambda} \right\} d\lambda = 0, \qquad (5.63)$$

where the functions $\varkappa_i (\lambda)$, $\omega_j (\lambda) \in W_2^{(1)}$, but in general in contrast to the Galerkin method ω_j do not coincide with the basis functions $\varkappa_j$.

We take $\varkappa_j$ to be piecewise linear functions equal to unity at j and to zero at the other points and choose ω_j as follows [5.14]:

$$\omega_j = \varkappa_j (\lambda) - F_{j+1/2} (\lambda),$$

$$\omega_{j+1} = \varkappa_{j+1} (\lambda) + F_{j+1/2} (\lambda), \quad jh \leqslant \lambda \leqslant (j+1) h,$$

where $F(\lambda) = a\lambda^2/h^2 + b\lambda/h + c$, with $F(\lambda)$ chosen so that it vanishes at the points $(j+1)h$, jh:

$$F_{j+1/2} (\lambda) = -a [(\lambda - jh)^2/h^2 - (\lambda - jh)/h], \quad jh \leqslant \lambda \leqslant (j+1) h;$$

then we have

$$\omega_j = \varkappa_j + a [(\lambda - jh)^2/h^2 - (\lambda - jh)/h],$$

$$\omega_{j+1} = \varkappa_{j+1} - a [(\lambda - jh)^2/h^2 - (\lambda - jh)/h], \qquad (5.64)$$

where

$$\varkappa_j = 1 - \lambda/h + j,$$

$$\varkappa_{j+1} = \lambda/h - j, \quad jh \leqslant \lambda \leqslant (j+1) h.$$

Now substituting (5.64) into (5.63) and assuming that the coefficients of (5.63) are step functions in every interval $(jh, (j+1)h)$ we get

$$\frac{d}{dt}\left\{\left(\frac{1}{6}-\frac{1}{12}d_{j-1/2}\right)a\cos\varphi H_{j-1/2}\psi_{j-1}\right.$$
$$+a\cos\varphi\left[\frac{2}{6}(H_{j-1/2}+H_{j+1/2})+\frac{1}{12}(d_{j-1/2}H_{j-1/2}-d_{j+1/2}H_{j+1/2})\right]\psi_j$$
$$\left.+\left(\frac{1}{6}+\frac{1}{12}d_{j+1/2}\right)a\cos\varphi H_{j+1/2}\psi_{j+1}\right\}$$
$$=\frac{1}{a\cos\varphi h^2}\left\{\left[\mu_{j+1/2}H_{j+1/2}+\frac{u_{j+1/2}H_{j+1/2}ha\cos\varphi}{2}\left(\frac{d_{j+1/2}}{3}-1\right)\right]\psi_{j+1}\right.$$
$$-[(\mu_{j+1/2}H_{j+1/2}+\mu_{j-1/2}H_{j-1/2})$$
$$+\frac{ha\cos\varphi}{2}\left(u_{j+1/2}H_{j+1/2}\frac{d_{j+1/2}}{3}+u_{j-1/2}H_{j-1/2}\frac{d_{j-1/2}}{3}\right)\Big]\psi_j$$
$$\left.+\left[\mu_{j-1/2}H_{j-1/2}+\frac{u_{j-1/2}H_{j-1/2}ha\cos\varphi}{2}\left(\frac{d_{j-1/2}}{3}+1\right)\right]\psi_{j-1}\right\}. \quad (5.65)$$

The family (5.65) of schemes dependent on the parameter d has some interesting properties. When $d = 0$, we have the usual finite-element scheme approximating the original equation (5.62) with the second order of accuracy with respect to the variable λ. Choosing the coefficient $d \neq 0$, sign (d) = sign (u) we arrive at a numerical scheme with additional dissipation. Of interest is scheme (5.65) with d chosen as

$$d=3\left[\cot h\left(\frac{uh\,a\cos\varphi}{2\mu}\right)-\frac{2\mu}{uh\,a\cos\varphi}\right]. \quad (5.66)$$

If $d = 3$ is chosen, then the first derivative with respect to λ is approximated by directed differences.

To solve (5.65) with respect to time we use either the Crank-Nicholson scheme or an implicit scheme (the natural filter scheme). Numerical solution schemes of the second and the third stage of coordinatewise splitting are similarly constructed. At each stage the problem is reduced to solving linear algebraic equations with a tridiagonal matrix which is inverted by factoring. We have thus given a complete description of a numerical algorithm for solving the problem of the diffusion-transport of oceanologic characteristics.

The system of equations for the adaptation of hydrological fields is as follows*:

$$H\frac{\partial u'}{\partial t}-H\left[l+(\tilde{u}+u')\frac{\tan\varphi}{a}\right]v'+\frac{1}{a\cos\varphi\rho_0}\left(I_\lambda\int_0^1 I_\lambda\,dz\right)$$
$$=-J_\lambda+\frac{H\tan\varphi}{a}\tilde{v}u',$$

*Here for ease of the presentation of the spatial approximation method instead of the equation for the integral stream function we use an equivalent problem [in terms of the planar components of the velocity vector.

$$H\frac{\partial v'}{\partial t}+H\left[l+(\tilde{u}+u')\frac{\tan\varphi}{a}\right]u'+\frac{1}{a\rho_0}\left(I_\varphi-\int_0^1 I_\varphi\,dz\right)$$
$$=-J_\varphi-\frac{H\tan\varphi}{a}\widetilde{uu'},$$

$$H\frac{\partial T}{\partial t}+\gamma_T w_2=0,$$

$$H\frac{\partial S}{\partial t}+\gamma_S w_2=0, \qquad (5.67)$$

$$H\frac{\partial\tilde{u}}{\partial t}-H\left[l+\frac{\tan\varphi}{a}\tilde{u}\right]\tilde{v}+\frac{H}{a\cos\varphi\,\rho_0}\frac{\partial p_0}{\partial\lambda}$$
$$=-\frac{1}{a\cos\varphi\,\rho_0}\int_0^1 I_\lambda\,dz+J_\lambda+\frac{\tau_1}{H},$$

$$H\frac{\partial\tilde{v}}{\partial t}+H\left[l+\frac{\tan\varphi}{a}\tilde{u}\right]\tilde{u}+\frac{H}{a\cos\varphi\,\rho_0}\frac{\partial p_0}{\partial\varphi}$$
$$=-\frac{1}{a\rho_0}\int_0^1 I_\varphi\,dz+J_\varphi+\frac{\tau_2}{H},$$

$$\frac{\partial H\tilde{u}}{\partial\lambda}+\frac{\partial}{\partial\varphi}(H\cos\varphi\,\tilde{v})=0.$$

We represent the dynamic space operator of problem (5.67) as a sum of two operators, corresponding to the left- and right-hand sides of equations (5.67). Carrying out an additional splitting within the adaptation stage we arrive at two systems of equations of which the more interesting (from the point of view of describing major large-scale oceanic motions) is the first, a linear system of equations of the form (5.67) with zero right-hand side.

We first consider the approximation of the three-dimensional problem: the first four equations for u', v', T, and S. The equations are approximated with respect to the space variables λ, φ, and z just as it has been done for the atmosphere dynamic model:

$$\frac{\partial u'_{i+1/2,\,j+1/2,\,k}}{\partial t}$$
$$-\left[l_{j+1/2}+\frac{\tilde{u}_{i+1/2,\,j+1/2}+u'_{i+1/2,\,j+1/2,\,k}}{a}\tan\varphi_{j+1/2}\right]v'_{i+1/2,\,j+1/2,\,k}$$
$$+\frac{1}{a\cos\varphi_{j+1/2}\rho_0}\left[p_k^h-\sum_{k=1}^{k}p_k^h\frac{\Delta z_{k+1/2}+\Delta z_{k-1/2}}{2}\right]=0,$$

$$\frac{\partial v'_{i+1/2,\,j+1/2,\,k}}{\partial t}$$

$$+\left[l_{j+1/2}+\frac{\tilde{u}_{i+1/2,\,j+1/2}+u'_{i+1/2,\,j+1/2,\,k}}{a}\tan\varphi_{j+1/2}\right]u'_{i+1/2,\,j+1/2,\,k}$$

$$+\frac{1}{a\rho_0}\left[Q_k^h-\sum_{k=1}^{k}\left[Q_k^h\frac{\Delta z_{k+1/2}+\Delta z_{k-1/2}}{2}\right]\right]=0, \qquad (5.68)$$

$$H_{i,\,j}\frac{\partial T_{i,\,j,\,k}}{\partial t}+\gamma_T w_2^h=0, \quad H_{i,\,j}\frac{\partial S_{i,\,j,\,k}}{\partial t}+\gamma_S w_2^h=0,$$

$$P_k^h=g\nabla_i\left[\left(\sum_{n=1}^{k}H\Delta z_{n-1/2}\,(\rho)_{n-1/2}\right)_{j+1/2}\right]$$

$$-gz_k\,((\rho_k)_{i+1/2})_{j+1/2}\,\nabla_i\,(H)_{j+1/2},$$

$$Q_k^h=g\nabla_j\left[\left(\sum_{n=1}^{k}H\Delta z_{n-1/2}\,(\rho)_{n-1/2}\right)_{j+1/2}\right]$$

$$-gz_k\,((\rho_k)_{i+1/2})_{j+1/2}\nabla_j\,(H)_{i+1/2}.$$

By introducing an integral stream function ψ using the relations

$$\frac{1}{H}\frac{\partial\psi}{\partial\varphi}=\tilde{u}, \quad \frac{1}{H\cos\varphi}\frac{\partial\psi}{\partial\lambda}=-\tilde{v}$$

the problem for the plane motion (the last three equations of (5.67) with the right-hand sides zero) is reduced to the following form:

$$-\frac{\partial}{\partial t}\left(\frac{\partial}{\partial\varphi}\frac{\cos\varphi}{H}\frac{\partial\psi}{\partial\varphi}+\frac{\partial}{\partial\lambda}\frac{1}{H\cos\varphi}\frac{\partial\psi}{\partial\lambda}\right)-\frac{\partial}{\partial\varphi}\frac{l}{H}\frac{\partial\psi}{\partial\lambda}$$

$$+\frac{\partial}{\partial\lambda}\frac{l}{H}\frac{\partial\psi}{\partial\varphi}=\frac{\partial Q_1}{\partial\varphi}-\frac{\partial Q_2}{\partial\lambda}, \qquad (5.69)$$

where

$$Q_1=\frac{\cos\varphi}{H\rho_0}\tau_1-\frac{g}{a\rho_0}\int_0^1\left(\frac{\partial}{\partial\lambda}H\int_0^z\rho\,dz-\frac{\partial H}{\partial\lambda}z\rho\right)dz,$$

$$Q_2=\frac{1}{H\rho_0}\tau_2-\frac{g}{a\rho_0}\int_0^1\left(\frac{\partial}{\partial\varphi}H\int_0^z\rho\,dz-\frac{\partial H}{\partial\varphi}z\rho\right)dz.$$

The peculiarity of the problem for an integral stream function is that it is required to be solved in a multiply connected region Ω_0 bounded by contours $\Gamma_0, \Gamma_1, \ldots, \Gamma_N$ which outline islands and individual continents. On one of the contours, on the outer contour Γ_0, for example, ψ may be set equal to zero; in accordance with the

no-flow condition ($u_n = 0$) on the remaining contours it is required to assume ψ constant. We thus have

$$\psi|_{\Gamma_0} = 0, \quad \psi|_{\Gamma_i} = C_i, \quad C_i = \text{const}, \quad i = 1, 2, \ldots, N. \tag{5.70}$$

To solve problem (5.69), (5.70) numerically one can use the projection- or the finite-difference scheme of the second order of approximation which reduce the problem to uniquely solvable systems of algebraic equtions [5.12]. A generalized solution of problem (5.69), (5.70) is the function

$$\overline{\psi}(t) \in \widetilde{W}_2^{(1)}$$

which satisfies for any $\omega \in \widetilde{W}_2^{(1)}$ at every instant of time t the integral relation

$$\frac{\partial}{\partial t}\int\limits_{\Omega_0}\left(\frac{\cos\varphi}{H}\frac{\partial\overline{\psi}}{\partial\varphi}\frac{\partial\omega}{\partial\varphi}+\frac{1}{H\cos\varphi}\frac{\partial\overline{\psi}}{\partial\lambda}\frac{\partial\omega}{\partial\lambda}\right)d\Omega_0$$

$$+\int\limits_{\Omega_0}\left(\frac{l}{H}\frac{\partial\overline{\psi}}{\partial\lambda}\frac{\partial\omega}{\partial\varphi}-\frac{l}{H}\frac{\partial\overline{\psi}}{\partial\varphi}\frac{\partial\omega}{\partial\lambda}\right)d\Omega_0$$

$$=-\int\limits_{\Omega_0}\left(Q_1\frac{\partial\omega}{\partial\varphi}-Q_2\frac{\partial\omega}{\partial\lambda}\right)d\Omega_0. \tag{5.71}$$

Here $\widetilde{W}_2^{(1)}$ is the space of functions which are square summable together with their first derivatives, equal to zero on Γ_0, and assume arbitrary constant values on Γ_i. To obtain a numerical scheme with respect to space variables one may either approximate (5.71) using discrete formulas or use the Galerkin process. In the latter case it is necessary to discretize Ω_0 in a certain way and choose functions with finite support to be basis functions by defining the M th approximation for $\overline{\psi}$ as follows:

$$\overline{\psi}^{(M)} = \sum_{i,\, j\in\Omega_0^h}\psi_{ij}\omega_{ij}+\sum_{k=1}^{N}C_k\overline{\omega}_k$$

Here $\overline{\omega}_k$ is a function equal to unity on the contour Γ_k and extended in some way (for example, linearly) to zero onto the remaining grid region and ω_{ij} is a function equal to unity at (i, j) and to zero at the other points.

5.5. THE SPLITTING METHOD BY PHYSICAL PROCESSES AND GEOMETRIC VARIABLES

As pointed out above, numerical integration of the equations of the dynamics of the ocean and atmosphere is based on the

splitting method [5.8], [5.9]. We describe the process of solving the problem in time using the example of the atmosphere dynamics model.

It was established in the preceding section that the dynamic operator of the problem, K, and its components have the skew symmetry property. This allows the Crank-Nicholson scheme to be used at each stage of splitting the system (5.61). Indeed we have

$$(\varphi^{j+1} - \varphi^j)/\tau + K(\varphi^{j+1} + \varphi^j)/2 = 0. \tag{5.72}$$

Performing a scalar multiplication of (5.72) by $(\varphi^{j+1} + \varphi^j)/2$ we get

$$(\varphi^{j+1}, \varphi^{j+1}) = (\varphi^j, \varphi^j). \tag{5.73}$$

Relation (5.73) implies the stability of the computing algorithm in the spectral norm $\| \varphi \|_2 = (\varphi, \varphi)^{1/2}$.

The operator K can be split into two operators, $K = K_T + K_A$, where K_T describes the process of the transport along the trajectories and K_A describes the adaptation of meteorological fields. This splitting of the dynamic operator K into operators describing different physical processes (splitting by physical processes) was first suggested by G. I. Marchuk. As we have established above, each of these operators has the skew symmetry property. The splitting scheme is of the form

$$\begin{aligned} &(\varphi^{j+1/2} - \varphi^j)/\tau + K_T(\varphi^{j+1/2} + \varphi^j)/2 = 0, \\ &(\varphi^{j+1} - \varphi^{j+1/2})/\tau + K_A(\varphi^{j+1} + \varphi^{j+1/2})/2 = 0. \end{aligned} \tag{5.74}$$

Scheme (5.74) has the second order of approximation in time provided K_T and K_A are commutative, and in the general case is of the first order of approximation, but by cyclic permutation it is possible to construct a scheme of the second order of approximation in the general case too. Since K_T and K_A are skew symmetric we have

$$\begin{aligned} &(\varphi^{j+1/2}, \varphi^{j+1/2}) = (\varphi^j, \varphi^j), \\ &(\varphi^{j+1}, \varphi^{j+1}) = (\varphi^{j+1/2}, \varphi^{j+1/2}). \end{aligned} \tag{5.75}$$

Hence

$$(\varphi^{j+1}, \varphi^{j+1}) = (\varphi^j, \varphi^j),$$

that is, in this case too the scheme is absolutely stable and the quadratic conservation laws are faithfully obeyed. As already shown, the operator K_T is a sum of three operators (the transport along each of the geometrical directions), each operator having the skew symmetry property. It is therefore natural that the problem should be split further. As a result we have the following algorithm:

$$\begin{aligned} &(\varphi^{j+1/8} - \varphi^j)/\tau + K_{T_1}(\varphi^{j+1/8} + \varphi^j)/2 = 0, \\ &(\varphi^{j+1/4} - \varphi^{j+1/8})/\tau + K_{T_2}(\varphi^{j+1/4} + \varphi^{j+1/8})/2 = 0, \end{aligned} \tag{5.76}$$

$$(\varphi^{j+1/2}-\varphi^{j+1/4})/\tau+K_{T_s}(\varphi^{j+1/2}+\varphi^{j+1/4})/2=0, \qquad (5.76)$$
$$(\varphi^{j+1}-\varphi^{j+1/2})/\tau+K_A(\varphi^{j+1}+\varphi^{j+1/2})/2=0.$$

That the difference scheme we have formulated is absolutely stable and that the quadratic functionals are exactly preserved can be established for (5.76) in a way similar to (5.75). Thus formally the algorithm has the first order of approximation with respect to time, the second order with respect to space variables, and is absolutely stable. In addition, a quadratic law of conservation (the conservation of total energy) exactly holds for it (under the assumptions made). The solution of the full diabatic problem is also carried out using the splitting method according to the following scheme (here we use the notation of the weak approximation method):

$\partial\varphi_1/\partial t+A_1\varphi_1=0,$ $\varphi_1(t_j)=\varphi(t_j)$	solving the equation of transport + diffusion
$\partial\varphi_2/\partial t+A_2\varphi_2=0,$ $\varphi_2(t_j)=\varphi_1(t_{j+1})$	solving the adaptation equations
$\partial\varphi_3/\partial t+A_3\varphi_3=0,$ $\varphi_3(t_j)=\varphi_2(t_{j+1})$	calculating the phase influxes of heat, dry and wet convection
$\partial\varphi_4/\partial t+A_4\varphi_4=0,$ $\varphi_4(t_j)=\varphi_3(t_{j+1})$	calculating radiation influx of heat
$\partial\varphi_5/\partial t+A_5\varphi_5=0,$ $\varphi_5(t_j)=\varphi_4(t_{j+1}).$	calculating the boundary layer, the dynamics of heat and soil moisture

5.6. NUMERICAL EXPERIMENTS IN SIMULATING PROCESSES OF A LARGE-SCALE CIRCULATION OF THE OCEAN AND ATMOSPHERE

The development of the software for the models was based on the idea of a software package by which was understood an oriented complex of interrelated application and system programs which ensure, on the basis of numerical simulation, solving problems related to studies of the general circulation of the ocean and atmosphere. In creating the package use was made of the modular principle consistently observing the following requirements [5.15]:

(i) modules should be developed independently and accumulated,

(ii) there should be "continuous dependence" on the model: small changes in the model should result in a small cost of the modification of the corresponding software complex;

(iii) for a given number of modules as many problems as possible should be implemented using them;

(iv) the efficiency of the computer resources used should be high.

The discrete model was constructed using the ideas of the splitting method. From the positions of the same method a structural analysis of the model was carried out which allowed each stage of the splitting to be formed as a closed module, its input and output parameters matched with the corresponding parameters of the other modules.

The programming language was FORTRAN. It was chosen because it is simple, versatile, allows the use of the modular principle, and has a developed routine library. To make the software as compatible with different computers as possible, several requirements were strictly observed in developing it. The principal requirements were the use of a limited alphabet of the FORTRAN language, with only the letters of the Latin alphabet, of all its letters, included, and the arrangement of exchange with external computer memory (magnetic drums, tapes, disks) using direct access files. The fulfilment of these requirements allowed the software package originally written for a BESM-6 computer to be readily adapted to computers of ES series, IBM and CRAY-1 computers.

The entire software complex is kept as a library of text modules on magnetic carriers*. Changes in the existing modules are made and the library is extended using a text editor. The latter's features allow models of various horizontal and vertical resolution to be constructed.

Now we discuss some numerical experiments we carried out on a BESM-6 computer at the Computing Centre of the Siberian Division of the USSR Academy of Sciences and on the CRAY-1 computer at the European Centre for Medium Range Weather Forecasts (Reading, Great Britain) in 1977 to 1981. The main aim of those experiments was to "reproduce" modern climate and to assess the validity of the model for solving problems in the study of climate and its changes. The calculations were carried out separately with the model of the general circulation of the atmosphere (the ocean surface temperature was fixed in time) and with the model of the general circulation of the ocean (the boundary conditions at the ocean surface given as known climatic distributions). In addition, integration was performed according to a general "ocean-atmosphere" model. The greater part of the experiments were carried out for the conditions of a

*The current version of the library contains about 10 000 FORTRAN operators, some 60% of which fall to the share of the atmospheric part of the model.

January mean circulation. The computations were made till the solutions reached the conditions of climatic equilibrium (except for the experiment dealing with the general oceanic and atmospheric circulation where no such equilibrium was reached). Choosing as initial conditions the state of rest and a given latitude- and longitude-independent temperature distribution (both in the ocean and the atmosphere), the time required for the atmosphere model to reach a quasi-equilibrium regime is about 60 model days and for the ocean model it is about 100 model years. An analysis of computational costs carried out on the CRAY-1 computer has shown that the processor time required by a computation according to the atmosphere model with a resolution of 5° in latitude and longitude and with three levels in the vertical amounts to 61 s for an integration for a period of twenty-four hours. For the ocean model with the same horizontal resolution and with eight levels in depth the processor amounts to 441 s time when integrating for a year.

Computations using global models of the general circulation of the ocean and atmosphere thus require a tremendous amount of computer time and therefore it becomes extremely important to map the algorithms used in them onto the architecture of computer systems with the aim of achieving a maximum speed when implemented on a computer. There are two different tendencies in the approach to this problem [5.16]: parallelizing computations using many functional devices (processors) and pipelining computations which involves increased loading of devices. The CRAY-1 computer belongs to machines with vector arithmetic (of the pipeline type). The structural analysis we have carried out has allowed us to identify effectively pipelinable algorithms (explicit iterative processes, streams of one-dimensional cyclic and scalar sweeps, data transfers into the memory, flows of problems involving the one-dimensional, in latitude, fast Fourier transform, etc.). "Pourly" written programs (ignoring the specific vector character of CRAY-1) considerably increase the processor time. Thus, for a variant of the software complex first written for BESM-6 and then adapted to CRAY-1, the time required to make a computation (with the above resolution) according to the atmosphere model for a day amounted to 331 s and according to the ocean model for a year amounted to 843 s.

Depending on the goals set the design, implementation, and analysis of numerical experiments were carried out along different lines. The first of them, which may be conveniently called "resolution" line included studies of the effect of the model discrete nature on the quasi-equilibrium climatic regime. Compared were cases of horizontal resolutions of 10° and 5° and of three and six levels in the vertical in the atmosphere model, and cases of resolution of 5° and of four and eight levels in the ocean model. As was to be expected, the improvement in the resolution results in a more realistic picture

of the general circulation of the atmosphere and large-scale motions in the ocean. With a further increase in the resolution necessary for a more exact reproduction of regional features of the circulation (such as the phenomena of blocking in the atmosphere or mesoscale eddies in the ocean) the ever growing difficulties of implementing numerical experiments force one to look at the problem from another point of view: to develop computer architecture to suit a particular problem rather than to fit software to a particular computer. The specific feature of the atmospheric circulation models is the parametric dependence of some physical processes (such as radiative transport and phase transitions of water) on horizontal coordinates. If enough processors are available, computations can be parallelized. Thus the development of hybrid concurrent-pipeline computer systems designed to simulate climate and its variation and produce long-term weather forecasts seems promising.

To illustrate the quality of the model, Figs. 5.2 and 5.3 present the temperature and geopotential fields at the 500 mb surface that were obtained in the experiment with a three-level resolution in the vertical and a five-degree resolution in latitude and longitude, Figs. 5.4 and 5.5 showing their deviation from the real climatic distributions for January [5.17]. On the whole the error of the reproduction of the climate temperature does not exceed 4 K in absolute value over the greater part of the terrestrial globe. At the same time there is a region located over East Eurasia and the Western Pacific at whose centre (about 40N, 140E) the error is as large as 12 K. In the same region the geopotential field error amounts to 160 dkm and, except the regions adjacent to Antarctica where, because of the absence of relief in the experiment, it may attain values of 240 dkm. In the remaining part of the globe the difference between the model geopotential field and its real climatic distribution for January is small ($\leqslant$80 dkm). An analysis has shown that the error structure of the temperature and geopotential fields, represented in Figs. 5.4 and 5.5, is independent of the model's resolution and is in fact its intrinsic property.

Figure 5.6 presents a chart of currents of the world ocean at a depth of 25 m and Fig. 5.7 presents for comparison a scheme of surface currents of many years' standing for winter in the Northern hemisphere [5.18]. It is easy to see that major large-scale circulation systems are well in evidence at a depth of 25 m in the Atlantic, Indian, and Pacific Oceans, as well as an Antarctic circumpolar current in the South Ocean. One can see a considerable divergence of currents near the equator in the Atlantic and Pacific Oceans and a rather weakened divergence in the Indian Ocean. This fact reflects the global upwelling in the equatorial belt, a distinctive feature of the circulation observed in the world ocean.

Another line in numerical experiments embraces studies of the

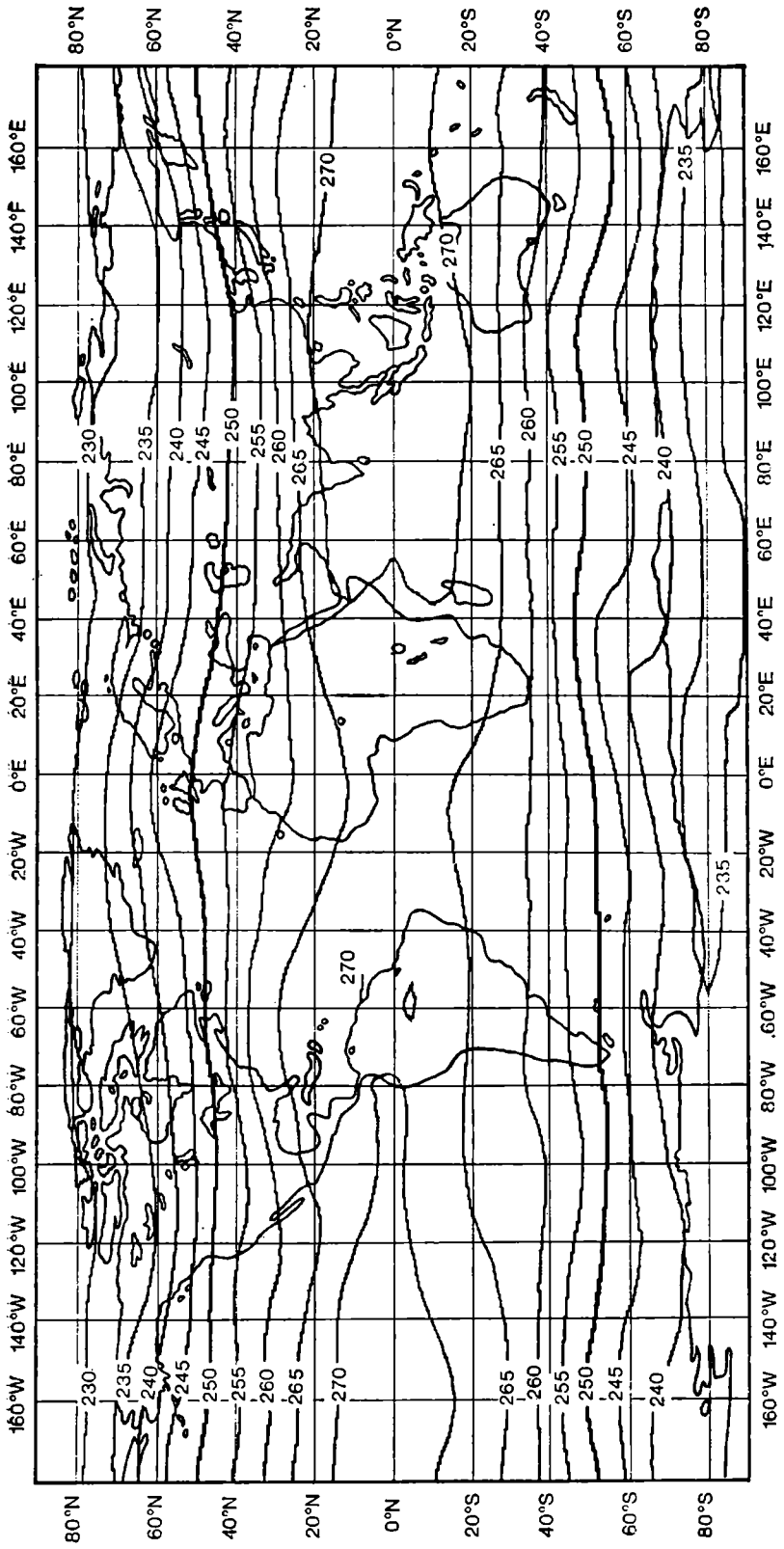


Fig. 5.2. Simulated temperature field at 500 mb surface

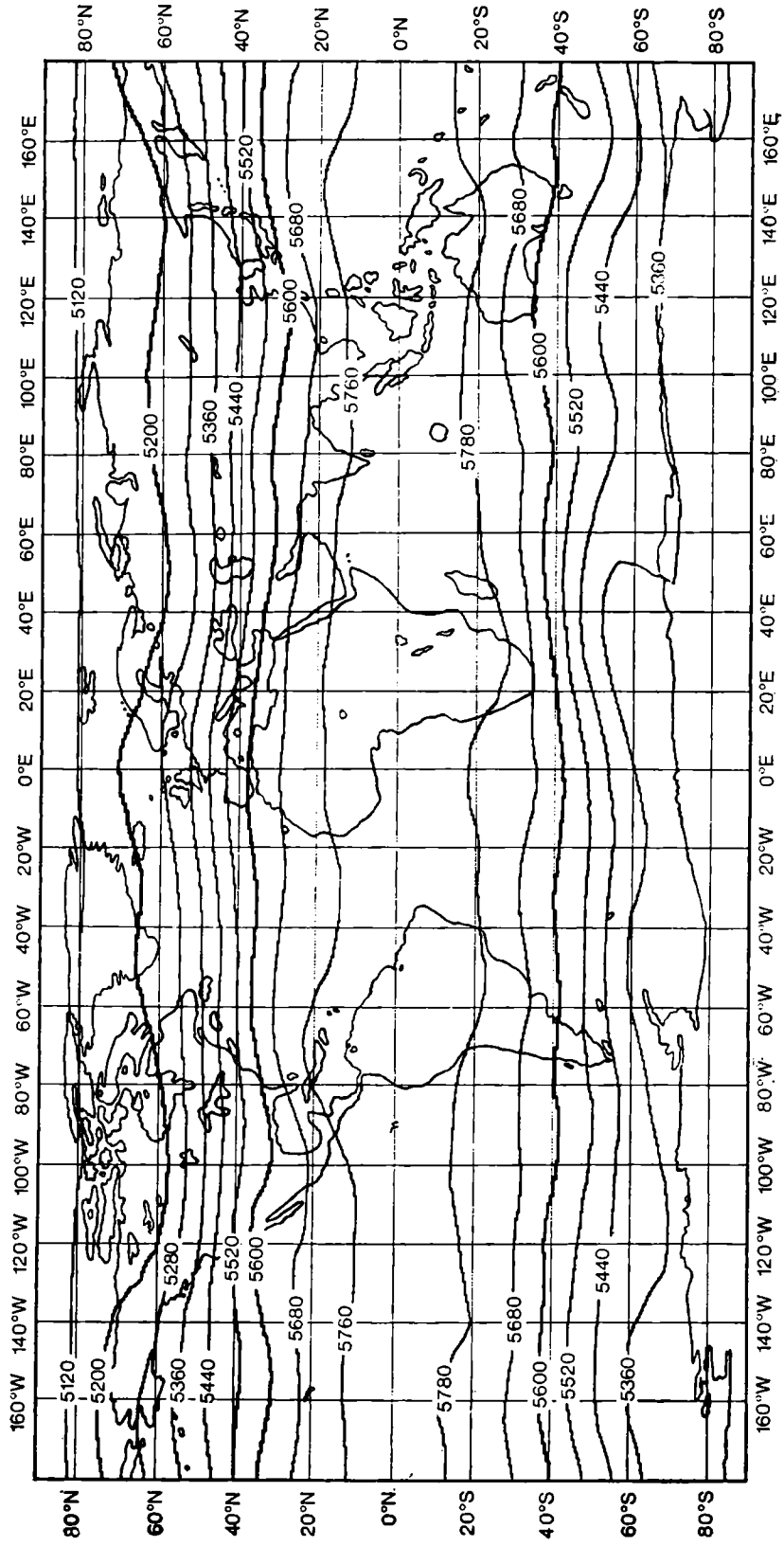


Fig. 5.3. Simulated geopotential field at 500 mb surface

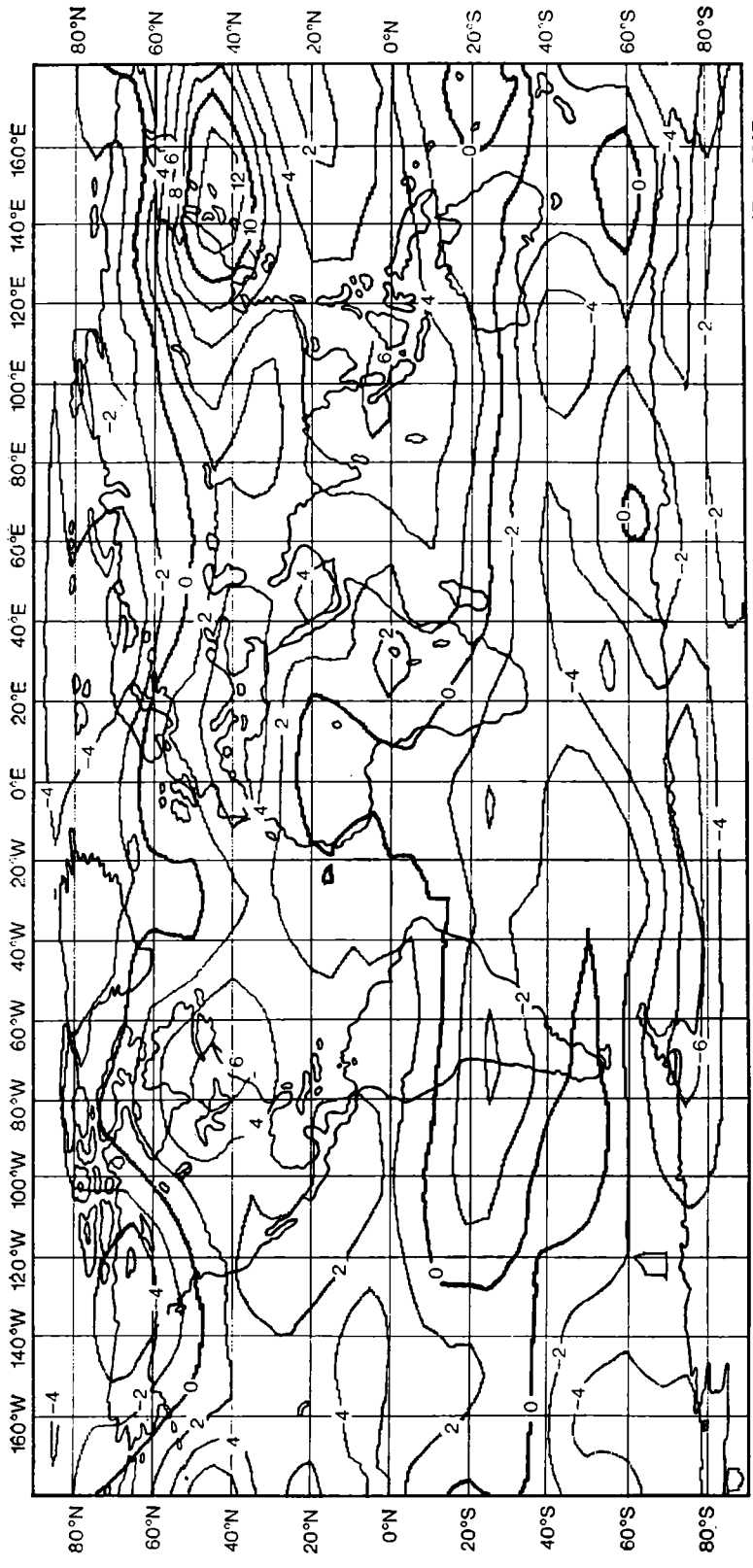


Fig. 5.4. Field of the difference of calculated and observed temperatures at 500 mb surface

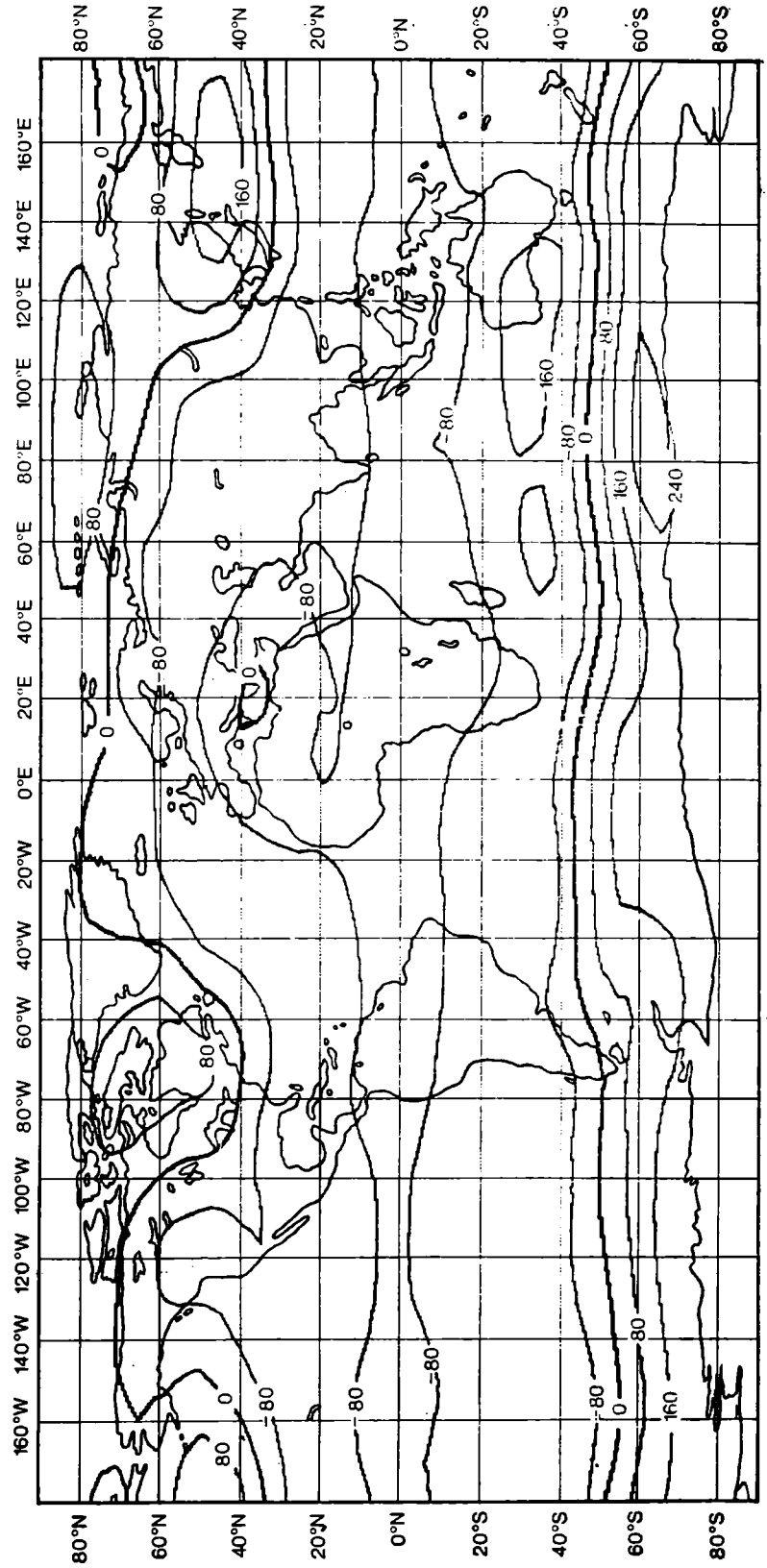


Fig. 5.5. Field of the difference of calculated and observed geopotentials at 500 mb surface

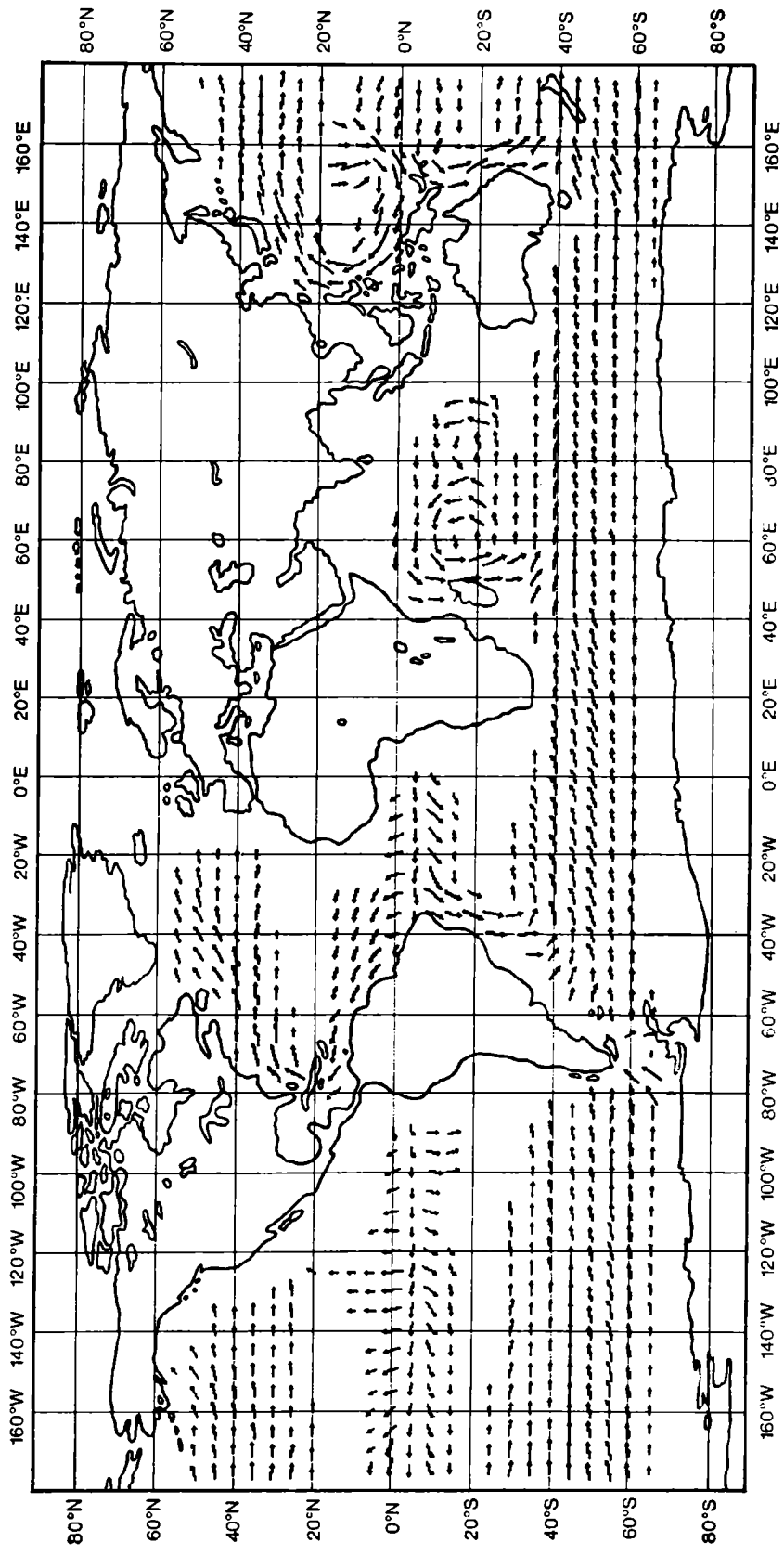


Fig. 5.6. Field of currents of the world ocean at a depth of 25 m

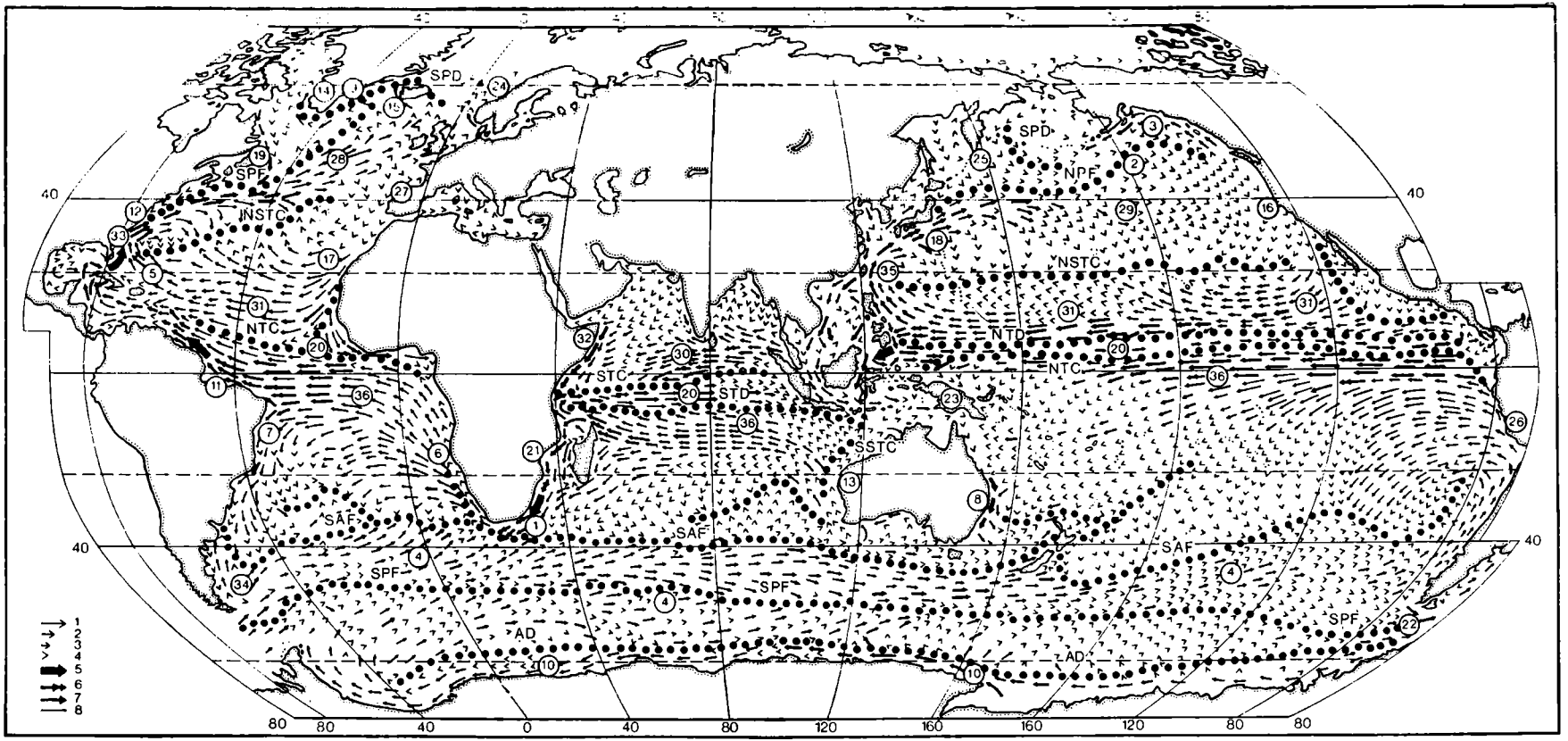


Fig. 5.7. Scheme of (winter) average surface flows of several years' standing in the world ocean

Fronts: NPF, North Polar; SAF, Subantarctic (Subtropical Convergence); SPF, South Polar (Antarctic Convergence). Divergences: SPD, Subpolar; NTD, North Tropical; STD, South Tropical; AD, Antarctic. Convergences: NSTC, North Subtropical; NTC, North Tropical; STC, South Tropical; SSTC, South Subtropical. Currents (figures in the circles): 1, Agulhas Current; 2, Aleutian Current; 3, Alaska Current; 4, Antarctic Circumpolar Current; 5, Antilles Current; 6, Bengali Current; 7, Brazilian Current; 8, East Australian Current; 9, East Greenland Current; 10, Western Winds Current; 11, Guianese Current; 12, Gulf Stream; 13, West Australian Current; 14, West Greenland Current; 15, Irminger Current; 16, Californian Current; 17, Canary Current; 18, Kuroshia; 19, Labrador Current; 20, Intertrade Counter-currents; 21, Mozambique Current; 22, Cap Horn Current; 23, New Guinea Current; 24, Norwegian Current; 25, Oyashia; 26, Peruvian Current; 27, Portuguese Current; 28, North Atlantic Current; 29, North Pacific Current; 30, North-western Monsoon Current; 31, Northern Trades Currents; 32, Somali Current; 33, Florida Current; 34, Falkland Current; 35, Formosan Current; 36, Southern Trades Currents. Current stability (characterized by the lengths of the arrows): 1, stable; 2, above the average; 3, below the average; 4, unstable. Current speed (characterized by the widths of the arrows), miles per day^{-1}: 5, over 36; 6, 24 to 36; 7, 12 to 24; 8, 0 to 12.

sensitivity of simulated climate with respect to changes in the various parametrizations used in the models. The results of these studies may be regarded from two points of view. On the one hand, it is the study of the degree of detailing required in constructing a parametrization of a particular physical process of subgrid scale. On the other hand, a particular physical problem is solved and it is necessary to have a clear idea of the physical mechanisms responsible for the process under study and of the degree of adequacy of the model to that particular physical process. The following problems in this direction were solved.

For the world ocean dynamics model we have estimated its sensitivity with respect to the macroturbulence exchange coefficients and investigated the role played by the nonlinear terms in the equations of motion. Using the atmospheric general circulation model the relation between the spatial and temporal kinetic energy spectrum was studied [5.4]. The solution of these problems is intimately related to the very structure of the computing algorithm which allows the problem of simulating the general circulation of the ocean and atmosphere to be solved with practically arbitrary coefficients in the terms that describe the dissipation of energy. In that sense a difference scheme permits investigating problems whose solution has an arbitrary "slope" of the amplitude-frequency characteristic in a given section of the spectrum.

Investigating the part played by cloud cover in the formation of climate and its variation is among the most important problems in modern climatology. To solve the problem of radiation-cloud relations a number of numerical experiments were conducted in which the coefficients in the interannual relations connecting the cloud cover and relative humidity were varied and different modification of modules describing radiative transport in the atmosphere were used. It was concluded from the analysis of these experiments that the problem of describing cloud cover is not a crucial one in the middle latitudes in winter since the clouds in that case are usually a mere indicator of the dynamics rather than its generator, the main sinks and sources generating the circulation in the middle latitudes in winter lying in the low and high latitudes [5.19].

The role played by a large-scale condensation in the formation of the basic characteristics of the atmospheric circulation in the middle latitudes of the Northern Hemisphere in winter time was studied. A comparative analysis of the results of two numerical experiments (with or without phase influxes of heat due to a large-scale condensation) has shown that on the whole the liberation of latent heat results in atmosphere stabilization. In other words, increasing the rate of built-up of baroclinically unstable waves in the presence of a large-scale condensation does not compensate fully for the increases in the average temperature due to the latent heat [5.20].

The sensitivity of the simulated January mean atmospheric general circulation with respect to the processes of turbulence interaction with the underlying surface was also studied. The following picture resulted from two numerical experiments which used two boundary-layer parametrization schemes differing mostly at low wind velocities (under the conditions characteristic of some regions of the tropical zone). An increased heat and moisture exchange of the tropical atmosphere with the ocean surface in the "abnormal" experiment as compared to the check one results in the intensification of the processes related to the phase transformations of moisture and in the increase of the total heating of the tropical atmosphere. This leads to substantial changes in the fundamental parameters of the climatic system (the zone average meridional and zonal temperature gradients and the static stability parameter) in the middle latitudes of the Northern Hemisphere and, as a consequence, to changes in the energy and dynamic characteristics of the circulation [5.21].

Belonging to a particular line important for the study of the interaction of the atmosphere and the ocean is the work dealing with the response of the atmosphere to the anomalies of the ocean surface temperature. Among the main difficulties that arise in tackling the problem by numerical experimentation with an atmospheric general circulation model is the discrimination of the signal against the noise due to the intrinsic variability of the model.*

The statistical significance of the result is usually determined by calculating at each node of the space grid the ratio $r = |\Delta|/\sigma$ ("signal"/"noise" ratio), where Δ is the deviation of climate ("anomaly-check") and σ^2 is the dispersion of sample climatic averages over the findings of check implementations. The parameter r is related to the t-Student criterion and the level of statistical significance. The latter's apriori level is given which corresponds to $r \geqslant 3$ and implies that the probability of the calculated variation of climate as randomly conditioned does not exceed 5% (with the number of check implementations not less than three).

Another approach is possible. Since the noise level in the middle latitudes in the model (and apparently in the atmosphere too) is due to the processes of hydrodynamic instability and to the nonlinear cascade of energy over the spectrum, it may be assumed that the spectral distribution of the signal-to-noise ratio is nonuniform, i.e. with a certain signal shape there may exist spectral windows where the ratio will be maximal. The experiment with the anomaly of the ocean surface temperature introduced in the area of the Icelandic

*The same problem also arises, of course, in the study of the model's sensitivity with respect to changes in the parametrizations of the processes of subgrid scale.

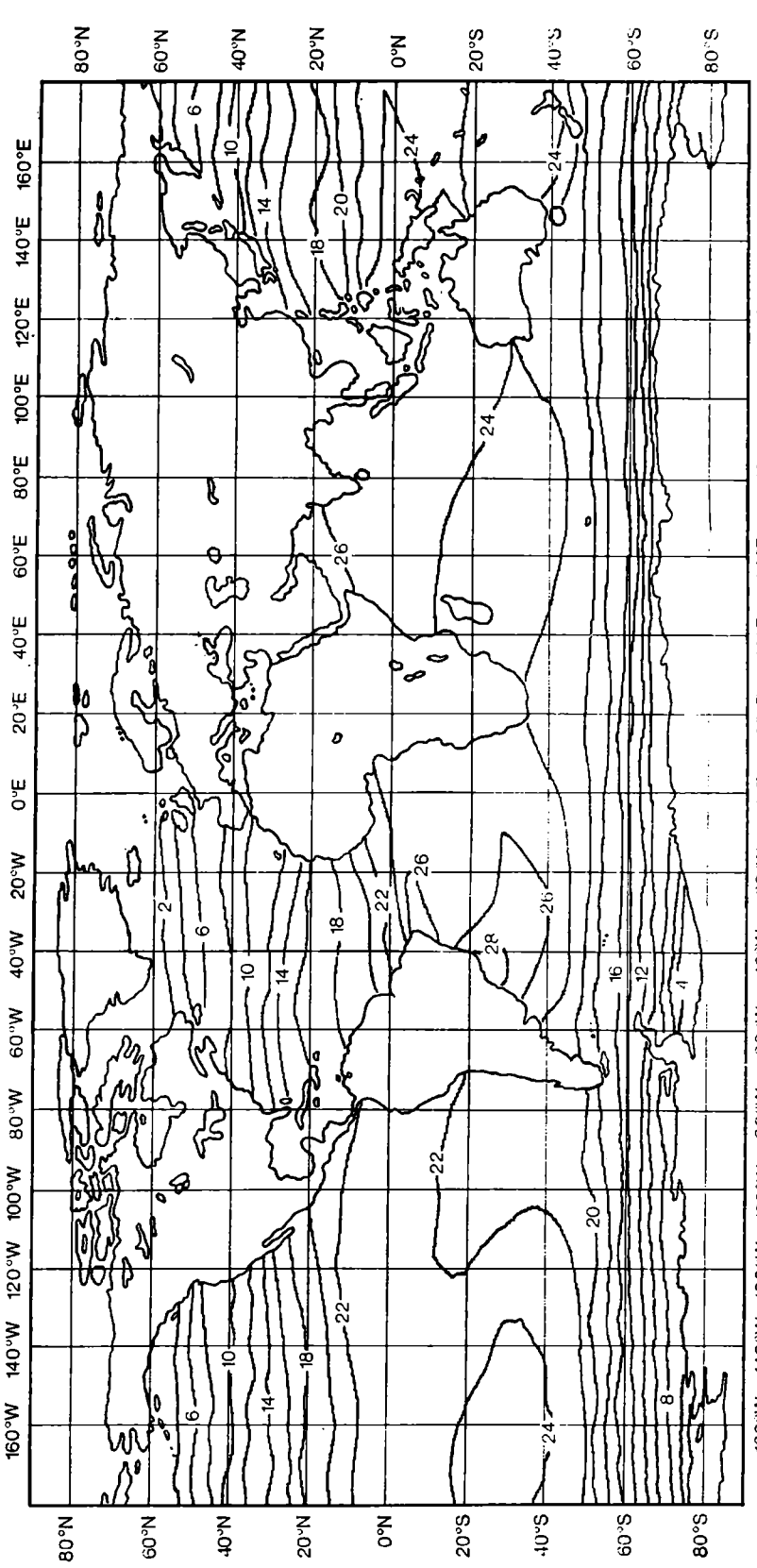


Fig. 5.8. Ocean surface temperature field calculated from the general "atmosphere-ocean" model

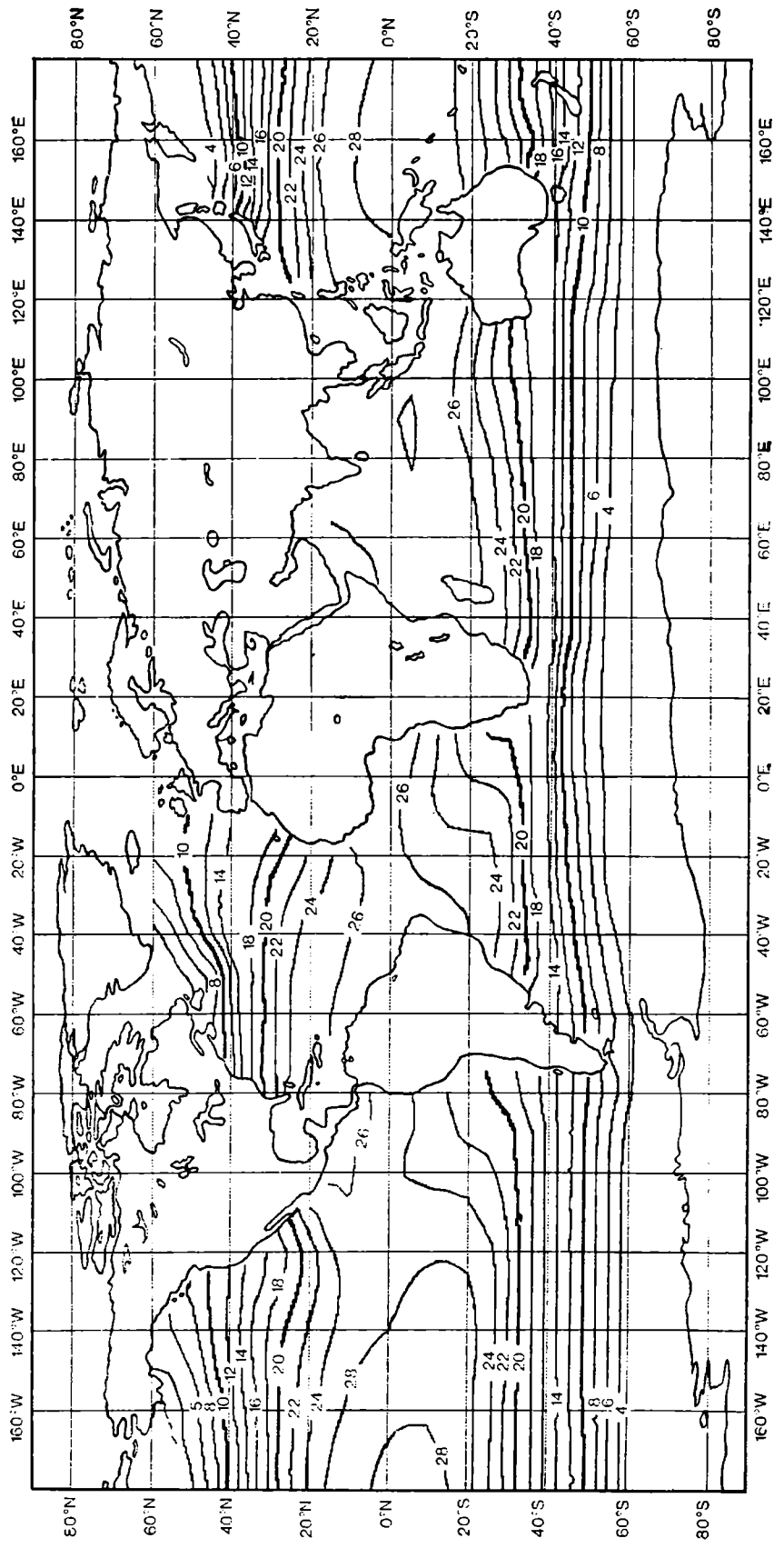


Fig. 5.9. Observed January mean ocean surface temperature distribution

Low has shown [5.22] that of statistical significance is the result for the zonal wave number 3 in the middle and high latitudes.

In conclusion we dwell on the problem of simulating the atmospheric-oceanic circulation. Mathematically, the fundamental question is how to construct a method for solving equations that describe the dynamics of the "atmosphere-ocean" system for long periods of integration with the purpose of reaching quasi-equilibrium conditions (if there are any). In our numerical experiment designed to simulate the January circulation of that system [5.2] we used the method proposed by Manabe and Bryan [5.23] and based on the artificial synchronization of characteristic time scales in the ocean and atmosphere. The experiment comprised two stages. First, with the January mean ocean surface temperature fixed, the atmospheric and the oceanic circulation were separately computed for periods corresponding to the quasi-equilibrium conditions in the ocean and atmosphere and then the general circulation was computed for a period of eleven "oceanic" years, which amounts to a term of about two months for the atmosphere. The quasi-equilibrium state for the joint system was not attained and appears to be unattainable in principle given the January mean declination of the Sun.

At the same time the experiment has led to a number of physical conclusions about the ocean-atmosphere interaction, in particular about the evolution of the ocean surface temperature. The ocean surface temperature field obtained at the end of the experiment (see Fig. 5.8) is in good qualitative agreement with observation (Fig. 5.9). It should be noted, however, that there are lower temperatures in the tropical zone of the Pacific and Indian Oceans (by about 3 °C to 4 °C) and higher temperatures off the shores of the Antarctica. In the former case this is probably due to a somewhat overestimated depth of the upper mixed layer and in the latter to the absence of a seasonal trend in the solar radiation. In the process of development in time, at the initial adaptation, of the atmospheric and oceanic parts of the system a distinct response is observed of the equatorial part of the ocean to the increase in the convective heat transfer into the atmosphere: a considerable intensification of the westward current, the intensification of the upwelling, and the corresponding deformation of the surface temperature, first the formation of a local minimum and then a decrease in its gradients owing to the intensified meridional circulation. Simultaneously the ejection of heat from the ocean rearranges the atmospheric circulation destroying cyclones off the west coast of the Pacific Ocean and in the North-Western Indian Ocean and thus removes the anomalies of heat flow from the ocean into the atmosphere.

To summarize, on the whole the worked out models reproduce sufficiently well the basic climatic characteristics of the actual ocean

and atmosphere and can be used to study the physics of climate and its variation.

References

5.1. G. I. Marchuk *et al.*, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 15, No. 5: 467-483 (1979) (in Russian).
5.2. G. I. Marchuk *et al.*, in: *Mathematical Modelling of the Dynamics of the Ocean and Atmosphere*, Part One (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980): 4-46 (in Russian).
5.3. G. I. Marchuk *et al.*, *GARP Publ.* 22: 318-349 (1979).
5.4. G. I. Marchuk, V. P. Dymnikov, and V. N. Lykosov, *ECMWF Tech. Memorandum*, No. 31 (1981).
5.5. N. A. Phillips, *J. Met.* 14, No. 2: 184-185 (1957).
5.6. O. I. Mamaev, *T, S-Analysis of Waters of the World Ocean* (Leningrad: Gidrometeoizdat, 1970) (in Russian).
5.7. V. N. Sinyaev, in: *Chisl. Metody Mekh. Sploshn. Sredy* (Novosibirsk: Vychisl. Tsentr SO AN SSSR 5, No. 2: 15-23 (1974) (in Russian).
5.8. G. I. Marchuk, *Numerical Methods in Weather Forecasting* (Leningrad: Gidrometeoizdat, 1967) (in Russian).
5.9. G. I. Marchuk, *Numerical Solution of Problems of the Dynamics of the Ocean and Atmosphere Based on the Splitting Method* (Leningrad: Gidrometeoizdat, 1974) (in Russian).
5.10. V. P. Dymnikov, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 14, No. 3: 243-250 (1978) (in Russian).
5.11. A. Arakawa, *Tech. Rep. No.* 7 (Dept. of Meteorol. Univ. of California, Los Angeles, 1972).
5.12. G. I. Marchuk, V. B. Zalesny, and V. I. Kuzin, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 11, No. 12: 1294-1300 (1975) (in Russian).
5.13. G. I. Marchuk, V. B. Zalesny, and V. N. Lykosov, *Simulating the Winter Climate of the World Ocean*, Preprint No. 32, 1982 (Moscow: OVM AN SSSR) (in Russian).
5.14. P. S. Huyakorn and K. Nilkuha, *Appl. Math. Modell.* 3, No. 1: 7-17 (1979).
5.15. N. I. Kozlov, *Organization of Computational Work* (Moscow: Nauka, 1981) (in Russian).
5.16. V. V. Voevodin, *A Mathematical Model of Pipeline Computations.* Preprint No. 42 (Moscow: AN SSSR, 1982) (in Russian).
5.17. R. L. Jenne, *NCAR Technical Note*, NCAR-TN/IA-111, 1975.
5.18. G. Schott, *Ann. Hydrograph. und marit Metheorol.* 71 (1943).
5.19. V. P. Dymnikov, V. Ya. Galin, V. N. Lykosov, *Report on the Third Session of the Joint Scientific Committee*, Dublin, 8-15 March (1982).
5.20. V. P. Dymnikov, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 18, No. 12: 1241-1246 (1982).
5.21. V. N. Lykosov and A. L. Kazakov, *Atmospheric Circulation Sensitivity to Processes of Microscale Turbulence Interaction with the Underlying Surface*, Preprint No. 459 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1983) (in Russian).
5.22. V. P. Dymnikov and V. N. Lykosov, *The Spectral Analysis of the Quasi-stationary Response of the Atmospheric Circulation to Anomalies of the Ocean Surface Temperature*, Preprint No. 61 (Moscow: OVM AN SSSR 1983) (in Russian).
5.23. S. Manabe and K. Bryan, *Mon. Weath. Rev.* 97, No. 11 (1969).

6 Methods of Mathematical Modelling of Sea and Oceanic Processes

A. S. Sarkisyan, V. P. Kochergin, V. I. Klimok, A. A. Kordzadze, V. I. Kuzin, V. A. Sukhorukov, and A. V. Shcherbakov

One of the promising methods of investigating processes occurring in the ocean is the method of mathematical modelling which allows, with efficient computing algorithms and powerful high-speed computers available, parameters of physical processes of interest to be determined with sufficient accuracy. This paper presents in a concise and fragmentary manner some results of investigations covering a wide range of problems of ocean dynamics.

6.1. CORRECTNESS AND UNIQUENESS OF SOLUTION IN MODELS OF OCEAN DYNAMICS

Before proceeding to the formulation and development of numerical algorithms for ocean dynamics problems it is necessary to discuss some very important questions, such as the study of the correctness of statements, uniqueness, existence, and stabilization of the solution, and obtaining the corresponding a priori estimates that turn out to be useful in constructing stable numerical schemes.

For linear ocean dynamics problems these questions were first identified and solved by Marchuk [6.1], Marchuk and Kagan [6.2], and Ovsyannikov [6.3]. A further development in this direction was the proof of the theorem of the uniqueness of solution by Kordzadze [6.4].

Bubnov and Kazhikhov [6.5] have established a unique solvability of linear nonstationary ocean dynamics problems.

Kochergin [6.6], [6.7] considered an ocean dynamics model in which the first two equations of motion contained no evolutionary and quasi-linear terms and the diffusion equation for the sea water density was written in full. The uniqueness of the classical solution has been proved for that model.

The existence of solutions of linear stationary and nonstationary and quasi-linear stationary sea current problems is proved in [6.8]-[6.10] using in the case of linear problems of explicit identification of barotropic components of currents proposed in [6.1], [6.11]. The solvability of problems for baroclinic components is investigated in [6.8]-[6.10], [6.12], [6.13]. In the case where the simplest (the equations of motion containing no turbulent friction forces) linear stationary and nonstationary models are considered the uniqueness

of solution is established and integral equations equivalent to boundary-value problems derived. Since the integral equations obtained are Fredholm equations of the second kind, the uniqueness of solutions of the boundary-value problems implies the solvability of equivalent integral equations and hence that of the boundary-value problems.

In the case of general linear boundary-value problems (with the equations of motion allowing for the turbulent friction forces), as in the preceding case, the corresponding a priori estimates are obtained and the solvability of the boundary-value problems considered is then proved using the results obtained earlier and the method of continuation with respect to the parameter.

In the study of quasi-linear stationary ocean models using the methods of the theory of monotonic operators a generalized solvability of these problems has been established (Kordzadze [6.10]).

Marchuk and Bubnov [6.11], [6.14], [6.15] have analyzed the main three-dimensional linearized models of the dynamics of a baroclinic ocean that allow for the oscillations of the level surface of the ocean with a variable bottom topography. Using the a priori estimates obtained the authors have established the correctness of the mathematical models. The a priori estimates obtained are necessary in the study of projection-difference schemes of solving quasi-stationary problems.

At present the most common method of obtaining steady-state solutions of ocean dynamics problems is the method of reaching the steady state in which a stationary or periodic solution is sought in the limit as $t \to \infty$ of the corresponding initial-boundary value problem. In [6.16] the method is substantiated for the main linearized models of the dynamics of a baroclinic ocean.

Existence and uniqueness theorems for problems of finding tides in outlying seas and oceanic basins are also proved by Marchuk and Kagan [6.2].

6.2. CONSTRUCTING NUMERICAL SCHEMES FOR THE OCEAN DYNAMICS EQUATIONS

6.2.1. The splitting method. The method is a powerful tool for obtaining in a simple and efficient form an approximate solution of many practical problems [6.1], [6.17]. Applying the splitting method in conjunction with well-known numerical methods often gives the latter a new interpretation and development.

The method proposed by Marchuk and Kuzin [6.18] and Kuzin [6.19], [6.20] in connection with calculating the global circulation in the world ocean and presented in Sec. 6.4 of the present paper is based on a combination of a finite-element scheme and the splitting

method, which results in a significantly simplified algorithm for obtaining grid equations and their solution. In the process of developing a model splitting is used at three different stages. The problem is first split by physical processes, the model reducing as a result to a collection of quasi-linear three-dimensional equations to be solved sequentially. The differential operators of these equations are then, after being transformed in a special way, broken down into a series of two-dimensional operators acting in sections that are parallel to (x, y), (x, z) and (y, z) planes. And, finally, after applying projection-difference methods the resulting grid operators are split into a series of one-dimensional operators. We give an example of this approach by considering the three-dimensional transport and diffusion equation for a substance, which is shown in Sec. 6.4 to be the basic element in the model of the world ocean based on the finite-element method. The substance transported may be the heat, salinity, pollutants, etc.

So consider the equation

$$\frac{\partial \varphi}{\partial t} + \operatorname{div}_3 \mathbf{u}\varphi - \sum_{i,\, j=1}^{3} \frac{\partial}{\partial x_i}\left(a_{i,j}\frac{\partial \varphi}{\partial x_j}\right) = 0, \quad a_{i,j} = a_{j,i}, \quad \varphi|_{t=0} = g, \tag{6.1}$$

in the region $Q = (0, T] \times G$, where G is a three-dimensional cylinder with lower base Ω_H, upper base Ω_0, lateral surface Σ, and height H. It is assumed that the continuity condition

$$\operatorname{div}_3 \mathbf{u} = 0 \tag{6.2}$$

holds. In equation (6.1) the last term describes turbulent diffusion. We transform (6.1) using a barotropic component of motion [6.1], and letting

$$u = \bar{u} + u', \quad v = \bar{v} + v', \quad w = w',$$
$$\bar{u} = (1/H)\int_0^H u\, dz, \quad \bar{v} = (1/H)\int_0^H v\, dz. \tag{6.3}$$

Relation (6.2) is split into two:

$$\operatorname{div}_2 \bar{\mathbf{u}} = 0, \quad \operatorname{div}_3 \mathbf{u}' = 0. \tag{6.4}$$

It is now possible to introduce the stream function ψ

$$\bar{u} = -(1/H)\partial\psi/\partial y, \quad \bar{v} = (1/H)\,\partial\psi/\partial x$$

and express the vertical velocity component in terms of u' and v':

$$w = \partial\hat{u}/\partial x + \partial\hat{v}/\partial y, \quad \hat{u} = \int_z^H u'\, dz, \quad \hat{v} = \int_z^H v'\, dz.$$

Then equation (6.1) can be written as

$$\frac{\partial \varphi}{\partial t} - \sum_{i,\,j=1}^{3} \frac{\partial}{\partial x_i}\left(A_{i,\,j}\frac{\partial \varphi}{\partial x_j}\right) = 0, \quad A_{i,\,j} \neq A_{j,\,i}, \tag{6.5}$$

where $A_{1,\,2} = a_{1,\,2} + \psi$, $A_{2,\,1} = a_{2,\,1} - \psi$, $A_{1,\,1} = a_{1,\,1}$, $A_{1,\,3} = a_{1,\,3} + \hat{u}$, $A_{3,\,1} = a_{3,\,1} - \hat{u}$, $A_{2,\,2} = a_{2,\,2}$, $A_{2,\,3} = a_{2,\,3} + \hat{v}$, $A_{3,\,2} = a_{3,\,2} - \hat{v}$, $A_{3,\,3} = a_{3,\,3}$. If we now represent the space operator as a sum of three two-dimensional operators

$$\Lambda = \Lambda_{1,2} + \Lambda_{1,3} + \Lambda_{2,3},$$

$$\Lambda_{1,2} = -\frac{1}{2}\frac{\partial}{\partial x_1}\left(A_{1,1}\frac{\partial}{\partial x_1}\right) - \frac{\partial}{\partial x_1}\left(A_{1,2}\frac{\partial}{\partial x_2}\right) - \frac{\partial}{\partial x_2}\left(A_{2,1}\frac{\partial}{\partial x_1}\right) - \frac{1}{2}\frac{\partial}{\partial x_2}\left(A_{2,2}\frac{\partial}{\partial x_2}\right),$$

$$\Lambda_{1,3} = -\frac{1}{2}\frac{\partial}{\partial x_1}\left(A_{1,1}\frac{\partial}{\partial x_1}\right) - \frac{\partial}{\partial x_1}\left(A_{1,3}\frac{\partial}{\partial x_3}\right) - \frac{\partial}{\partial x_3}\left(A_{3,1}\frac{\partial}{\partial x_1}\right) - \frac{1}{2}\left(\frac{\partial}{\partial x_3}A_{3,3}\frac{\partial}{\partial x_3}\right),$$

$$\Lambda_{2,3} = -\frac{1}{2}\frac{\partial}{\partial x_2}\left(A_{2,2}\frac{\partial}{\partial x_2}\right) - \frac{\partial}{\partial x_2}\left(A_{2,3}\frac{\partial}{\partial x_3}\right) - \frac{\partial}{\partial x_3}\left(A_{3,2}\frac{\partial}{\partial x_2}\right) - \frac{1}{2}\left(\frac{\partial}{\partial x_3}A_{3,3}\frac{\partial}{\partial x_3}\right),$$

then it is possible to show that for the coefficients $a_{i,\,j}$ appearing in the ocean models (see Sec. 6.4) all the operators $\Lambda_{i,\,j}$ are positive semidefinite. This makes it possible to use the splitting method in solving equation (6.5). At each stage a series of problems is solved on sections that are parallel to the coordinate planes. To discretize the two-dimensional problems obtained according to space we use the projection-difference method based on piecewise linear completions which leads to balance approximations similar in quality to the well-known Arakawa schemes.

We discuss briefly solving two-dimensional problems. The method of approximating problems of this kind was developed by Kuzin [6.19], Marchuk *et al.* [6.21], Marchuk and Kuzin [6.18]. Consider the equation

$$\partial\varphi/\partial t + \Lambda\varphi = f, \quad \varphi|_{t=0} = g,$$

$$\Lambda = -\sum_{i,\,j=1}^{2} \frac{\partial}{\partial x_i}\left(A_{i,\,j}\frac{\partial}{\partial x_j}\right), \quad A_{i,\,j} \neq A_{j,\,i}, \tag{6.6}$$

in a two-dimensional region Ω. The boundary conditions are not specified (these may be Dirichlet conditions or generalized Neumann

conditions). The grid equations are constructed using the Galerkin method. To define the coordinate functions a grid region Ω^h with spacings in coordinate directions h and k is introduced. Rectangular cells δ_{mn} of the grid Ω^h are then triangulated by diagonals of positive or negative slope depending on the sign of the integral

$$\int_{\delta_{mn}} [(A_{1,2} + A_{2,1})/2]\, d\Omega.$$

Then a piecewise linear approximate solution is sought on the triangulation. Trial functions are also piecewise linear continuous functions

$$\bar{\varphi}(x_1, x_2, t) = \sum_{m,n} \varphi_{m,n}(t)\, w_{m,n}(x_1, x_2). \tag{6.7}$$

The coefficients $\varphi_{m,n}(t)$ are found from the system of Galerkin equations

$$\begin{aligned} &\frac{\partial \varphi_{m,n}}{\partial t}(1, w_{m,n}) + J(\bar{\varphi}, w_{m,n}) = (f, w_{m,n}) \\ &(u, v) = \int_{\Omega} uv\, d\Omega, \quad J(u, v) = (\Lambda u, v) \end{aligned} \tag{6.8}$$

which in operation representation becomes

$$M^h\, \partial\Phi/\partial t + \Lambda^h \Phi = F. \tag{6.9}$$

Here M^h is the positive definite operator and $\Lambda^h \geqslant 0$ is the grid operator approximating the space operator of the equation Λ. A detailed analysis of the structure of the operator Λ^h shows [6.18] that it may be represented as a sum of four three-point operators

$$\Lambda^h = \Lambda^h_{1,1} + \Lambda^h_{2,2} + \Lambda^h_{1,2} + \Lambda^h_{2,1}$$

acting in two coordinates $(\Lambda^h_{1,1}, \Lambda^h_{2,2})$ and two diagonal directions $(\Lambda^h_{1,2}, \Lambda^h_{2,1})$. It can also be shown that one-dimensional operators are positive semidefinite if the conditions

$$\begin{aligned} A_{1,1} &\geqslant (h/k)\, |\, (A_{1,2} + A_{2,1})/2\, |, \\ A_{2,2} &\geqslant (k/h)\, |\, (A_{1,2} + A_{2,1})/2\, | \end{aligned} \tag{6.10}$$

hold. Conditions (6.10) are stronger than the conditions of uniform ellipticity of the original operator. The positive semidefiniteness of the one-dimensional grid operators $\Lambda^h_{i,j}$ makes it possible to use splitting methods for solving equation (6.9) with respect to time, in particular the two-cycle method ensuring a second order of accuracy. Implementation of the entire algorithm is effected by a series of sweeps.

Now we proceed to describe the general structure of the solution of the system of ocean dynamics equations [6.2], [6.21]-[6.23]. In

operator form the system becomes

$$B\,\partial\varphi/\partial t + A\varphi = 0 \tag{6.11}$$

with the initial conditions at $t = 0$

$$B\varphi = BF, \tag{6.12}$$

where B and A are the matrix operators corresponding to the system of equations.

For a numerical solution of the problem we consider the time interval $0 \leqslant t \leqslant T$ to be broken down into equal intervals $t_{j-1} \leqslant t \leqslant t_j$ with a time step $\tau = t_j - t_{j-1}$ and to solve problem (6.11), (6.12) apply in each extended interval $t_{j-1} \leqslant t \leqslant t_{j+1}$ the two-cycle splitting method [6.1], [6.17]. Assuming that φ^{j-1} is the solution at t_{j-1}, we then have

$$B\,\partial\varphi_1/\partial t + A_1\varphi_1 = 0, \quad B\varphi_1^{j-1} = B\varphi^{j-1} \tag{6.13}$$

in the first interval $t_{j-1} \leqslant t \leqslant t_j$ and

$$B\,\partial\varphi_2/\partial t + A_2\varphi_2 = 0, \quad B\varphi_2^{j-1} = B\varphi_1^{j} \tag{6.14}$$

in the interval $t_{j-1} \leqslant t \leqslant t_{j+1}$, and in the last interval $t_j \leqslant t \leqslant t_{j+1}$ we solve the probem

$$B\,\partial\varphi_3/\partial t + A_1\varphi_3 = 0, \quad B\varphi_3^{j} = B\varphi_2^{j+1}, \tag{6.15}$$

where A_1 is the operator containing transport and diffusion terms and A_2 is the operator of dynamic adaptation of hydrophysical fields. Note that A_1 and A_2 add up to the operator A of (6.11) and that they are nonnegative and skew-symmetric operators, respectively [6.17].

Let Ω be a region covered with a grid Ω^h uniform in x and y and nonuniform in z. Then at the first and third stages of splitting consider for problems (6.13) and (6.15) the difference approximation [6.1], [6.22] represented diagrammatically in Fig. 6.1. As a result we arrive at a scheme of the second order of accuracy in x and y and the first order of accuracy in z. Then in the intervals $t_{j-1} \leqslant t \leqslant t_j$ and $t_j \leqslant t \leqslant t_{j+1}$, applying the two-cycle componentwise splitting method we arrive at a collection of one-dimensional problems that can be efficiently implemented by factoring. The Crank-Nicholson scheme is used at each elementary step. Before using the splitting method the barotropic component $\overline{\varphi}$ and baroclinic component φ' of the solution of the problem φ are separated out.

We apply the Crank-Nicholson scheme to the problem for the barotropic component $\overline{\varphi}$, introduce the difference stream function, and after appropriate transformations arrive at a difference equation for the stream function [6.22] which can be solved by various direct and iterative methods.

Consider next the solution of the problem for the baroclinic components of the solution φ'. To this end the operator corresponding to a given system of equations is represented as a sum of the operators A_l, A_{xOz} and A_{yOz}. As a result we obtain a collection of simpler problems split according to the sections xOz and yOz and a system

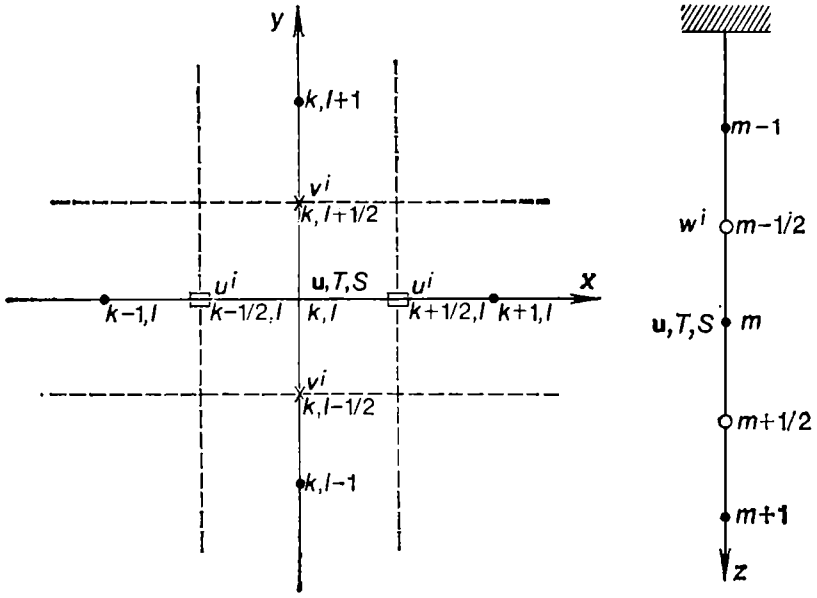


Fig. 6.1. Scheme of a difference approximation

including only terms with a Coriolis force. Since these operators are not commutative, we use the two-cycle splitting method. To solve the problems on the sections we introduce the difference stream function, and as a result of appropriate transformations the problems reduce to the solution of difference equations similar in structure to the equations for the barotropic component.

6.2.2. Increasing the order of approximation by the method of nested grids. Differential equations with a small parameter of the higher derivative occur in many problems of applied mathematics. In dealing with oceanic currents these are problems of finding the integral stream function or the level surface. Of the same type are the three-dimensional linearized heat and salinity transport equations (and the density equation). Similar problems are solved in viscous fluid dynamics when flows with large Reynolds numbers are calculated.

Consider thus in a two-dimensional region Ω_0 with boundary Γ_0 the elliptic equation

$$\mathcal{L}\eta \equiv -\varepsilon \Delta\eta + A\, \partial\eta/\partial x + B\, \partial\eta/\partial y = f \qquad (6.16)$$

with boundary conditions

$$\eta\, \Gamma_0 = \psi^*. \qquad (6.17)$$

Here $A(x, y)$, $B(x, y)$, $f(x, y)$, $\psi^*(x, y)$ are known sufficiently smooth functions and ε is a small parameter such that

$$\varepsilon \ll \max_{\Omega_0} [|A(x, y)|, \quad |B(x, y)|]. \tag{6.18}$$

Assuming for simplicity the region Ω_0 to be a unit square we introduce a uniform grid Ω_0^h with nodes (x_j, y_i), $x_j = hj$, $y_i = hi$, $h = \Delta x = \Delta y$. If equation (6.16) is approximated by directed differences, then the scheme viscosity is known to exceed significantly the physical viscosity for large grid spacings. But if the central-difference scheme is used it will describe correctly the solution of the equation only for grid spacings $h < h_0$

$$h_0 = 2\varepsilon/\max_{\Omega_0} (|A|, |B|). \tag{6.19}$$

If these conditions are violated, the difference solution will give strong oscillations. A similar condition is also required for other schemes with symmetric approximation of the first derivatives of equation (6.16). (The solution of problem (6.16) is carried out by the method suggested in [6.1].)

Shaidurov [6.24] has developed a method of increasing the order of approximation on a sequence of nested grids. The idea behind this method is due to Richardson, but the method was first applied and justified by Volkov [6.25] for the Laplace equation. The essence of the method is to find several difference solutions on different grids, after which a linear combination of those solutions allows the accuracy of calculations to be significantly increased. For details of this method see Marchuk and Shaidurov [6.26].

The method of a grid sequence is very efficient for solving equation (6.16) (Kochergin and Shcherbakov [6.27]). Along with Ω_0^h we introduce a nested grid Ω_0^{2h} each node of which (x_j, y_i), $i, j = 2, 4, \ldots, n-2$, is a node of Ω_0^h. On both grids we use a scheme of directed differences, i.e. we solve two problems:

$$(\Lambda^h + C^h)\,\eta^h = f^h, \quad \eta^h = \psi^* \quad \text{on} \quad \Gamma_0^h, \tag{6.20}$$

$$(\Lambda^{2h} + C^{2h})\,\eta^{2h} = f^{2h}, \quad \eta^{2h} = \psi^* \quad \text{on} \quad \Gamma_0^{2h}. \tag{6.21}$$

Here Λ^{kh} is the operator of the central-difference scheme and C^{kh} is the operator approximating the scheme viscosity. On finding η^h from (6.20) and η^{2h} from (6.21) we find $\tilde{\eta}^{2h}$ on Ω_0^{2h} from the formula

$$\tilde{\eta}^{2h} = 2\eta^h - \eta^{2h} \tag{6.22}$$

which has a second order of accuracy [6.26], [6.28].

The method of increasing the order of approximation of difference schemes with the aid of a sequence of nested grids has turned out to be efficient in solving the mixed boundary-value problem (Kochergin and Shcherbakov [6.27]) and the Poincaré problem (Kochergin, Klimok, and Shcherbakov [6.29]). The efficiency of the nested-grid

method is somewhat affected by having to interpolate the solution obtained on the nested grid Ω_0^{2h} to the basic grid Ω_0^h. To avoid the indicated difficulties and to deal with a single difference grid it is suggested in [6.27] that the scheme viscosity should be eliminated by the method of successive approximation. This leads to a one-parameter family (with parameter α) of nine-point schemes of the second order of approximation for problem (6.16) to (6.19)

$$\frac{4+6\alpha}{3}(\Lambda^h + C^h)\,\eta^h = f^h + \left(\frac{1+6\alpha}{3}\,\Lambda^{2h} + \frac{2+3\alpha}{3}\,C^{2h}\right)\eta^h. \tag{6.23}$$

The optimum value of the parameter $\alpha = 1/6$ has been found by solving the model equations.

6.3. VERTICAL TURBULENT EXCHANGE

The model of vertical turbulent exchange proposed at the Computing Centre of the Siberian Division of the USSR Academy of Sciences [6.30] is based on the dynamic equations of turbulence: those of turbulence energy b and turbulent dissipation velocity ε. The turbulent transport coefficient k is given by the simple algebraic relation

$$k = \text{const} \cdot b^2/\varepsilon. \tag{6.24}$$

The semiempirical approach under discussion was sketched in a short note by Kolmogorov [6.31] (and still earlier by Zagustin [6.32]) where turbulence was characterized by two principal parameters, energy and scale. The idea of using the dissipation equation for the closure of the Fridman-Keller equations was first put forward by Chou [6.33] and then it was taken up again in the late 1950's by Davydov [6.34]. A further development of the idea was displayed in the work of American [6.35], British [6.36], and other investigators.

We write the dynamic equations of turbulence in the boundary-layer approximation for problems of geophysical hydrodynamics [6.30]

$$b_t = k\,(u_z^2 + v_z^2) + (kb_z)_z - \varepsilon - (g/\rho_0)\,k\rho_z, \tag{6.25}$$

$$\varepsilon_t = 1.38\,(\varepsilon/b)\,k\,(u_z^2 + v_z^2) + (k\varepsilon_z)_z - 1.4\,(\varepsilon/b)\,\varepsilon - 1.4\,(g/\rho_0)\,(\varepsilon/b)\,k\rho_z. \tag{6.26}$$

Here the indices designate differentiation with respect to the corresponding variables, u and v are the projections of velocity onto the horizontal coordinate axes, ρ is the density, z is the vertical coordinate directed to the Earth's centre, and g is the acceleration due to gravity.

Equations (6.24) to (6.26) supplemented with the equations of

hydrothermodynamics are intended for the study of the dynamics of turbulent processes in a stratified medium. The empirical constants in the equations must generally be functions of dimensionless parameters of the problem. In finding these constants the laws of the degeneration of isotropic turbulence, the relations of the logarithmic boundary layer and the results of laboratory experiments were used in [6.37].

The system of turbulence equations (6.24) to (6.26) was approved on ocean dynamics problems. The investigation of oceanological problems has shown that the large-scale structure of upper quasi-homogeneous layer (UQL) can be described using the Kolmogorov-Obukhov formulas [6.38]

$$k = L^2 [u_z^2 + v_z^2 - (g/\rho_0)\, \rho_z]^{1/2} \tag{6.27}$$

(here the indices again designate differentiation with respect to the corresponding variables). The path of mixing L is proportional to the thickness of the turbulent layer, $L = \text{const} \cdot h$. Formula (6.27) is a consequence of equations (6.24) to (6.26) if one ignores in (6.25) the nonstationary and diffusion terms and uses the Kolmogorov "law of 3/2", $\varepsilon = \text{const} \cdot b^{3/2}/l$ (l is a correlation scale).

A study of oceanological problems using equations (6.24) to (6.26) has shown that in the system of turbulence equations the diffusion terms become important at the lower boundary of the turbulent region in the temperature jump zone [6.39]. In that region the turbulence scale l can be approximated with a linear depth function by putting $l \sim 1 - z/h$, an analogy with a boundary layer at a rigid surface. Note that the value of the constant $c = 0.05$ obtained in finding the mixing path length $L = ch$ agrees well with the known data on the scale of turbulent formations in the UQL.

The dynamic turbulence equations (6.24) to (6.26) are more meaningful than formula (6.27). They make it possible, for example, to define on the temperature jump the boundary layer which forms during a storm [6.30]. Formula (6.27) does not describe this phenomenon, nor is it possible to obtain a smooth steady-state profile of the coefficient k with improved resolution ($\Delta z < 0.5$ m) in the temperature jump zone. This appears to be due to the approximate assumption that the mixing path length L is proportional to h regardless of the depth.

An analysis of the numerical results for the UQL has made it possible to obtain analytical solutions vividly demonstrating the role played by the diagnostic external variables in the formation of the UQL [6.40].

6.4. NUMERICAL SIMULATION OF WORLD OCEAN CIRCULATIONS

We illustrate how to use the numerical methods of Sec. 6.2 in constructing particular ocean dynamics models by way

of calculations of hydrodynamical characteristics of the world ocean carried out at the Computing Centre of the Siberian Division of the USSR Academy of Sciences, using three different models which differ in the details of physical processes and the numerical methods used.

To begin with, we write a system of ocean hydrothermodynamics equations in the most general form and indicate in each particular case what approximations are made. In spherical coordinates the system is of the form

$$\frac{du}{dt}+\frac{mn}{a}uv+2\omega nv=-\frac{m}{a\rho_0}\frac{\partial p}{\partial\lambda}+\frac{\partial}{\partial z}\left(k\frac{\partial u}{\partial z}\right)+F^{\lambda}-Ru, \tag{6.28}$$

$$\frac{dv}{dt}-\frac{mn}{a}uu-2\omega nu=-\frac{1}{a\rho_0}\frac{\partial p}{\partial\theta}+\frac{\partial}{\partial z}\left(k\frac{\partial v}{\partial z}\right)+F^{\theta}-Rv, \tag{6.29}$$

$$\frac{\partial p}{\partial z}=g\rho, \tag{6.30}$$

$$\frac{\partial w}{\partial z}+\frac{m}{a}\left(\frac{\partial u}{\partial\lambda}+\frac{\partial}{\partial\theta}\left(\frac{v}{m}\right)\right)=0, \tag{6.31}$$

$$\frac{dT}{dt}=\frac{\partial}{\partial z}\left(k_T\frac{\partial T}{\partial z}\right)+\frac{A_H}{a^2}\Delta T, \tag{6.32}$$

$$\frac{dS}{dt}=\frac{\partial}{\partial z}k_S\frac{\partial S}{\partial z}+\frac{A_H}{a^2}\Delta S, \tag{6.33}$$

$$\rho=f(T,\ S). \tag{6.34}$$

The boundary conditions common to all the models are of the form

$$z=0:\ k\rho_0\frac{\partial u}{\partial z}=-\tau_{\lambda},\ k\rho_0\frac{\partial v}{\partial z}=-\tau_{\theta},\ w=0,\ T=T^*,\ S=S^* \tag{6.35}$$

$$z=H:\ w=\frac{mu}{a}\frac{\partial H}{\partial\lambda}+\frac{v}{a}\frac{\partial H}{\partial\theta}, \tag{6.36}$$

$$t=0:\ u=u^0,\quad v=v^0,\quad T=T^0,\quad S=S^0. \tag{6.37}$$

Here

$$\frac{d\varphi}{dt}=\frac{m}{a}\left(\frac{\partial(u\varphi)}{\partial\lambda}+\frac{\partial[(v/m)\varphi]}{\partial\theta}\right)+\frac{\partial(w\varphi)}{\partial z},$$

$$F^{\lambda}=\frac{A_H}{a^2}\left[\Delta u+(1-m^2n^2)u+2m^2n\frac{\partial v}{\partial\lambda}\right],$$

$$F^{\theta}=\frac{A_H}{a^2}\left[\Delta v+(1-m^2n^2)v-2m^2n\frac{\partial u}{\partial\lambda}\right],$$

$$\Lambda\varphi=m^2\left[\frac{\partial^2\varphi}{\partial\lambda^2}+\frac{1}{m}\frac{\partial}{\partial\theta}\left(\frac{1}{m}\frac{\partial\varphi}{\partial\theta}\right)\right],\quad n=\cos\theta,\quad m=1/\sin\theta.$$

Among the boundary conditions (6.35) and (6.36) there are no conditions on the lateral boundary Γ and, with $z=H$, there are none for functions u, v, T and S. They will be specified for each particular case.

6.4.1. The finite-element model. This model incorporates equations (6.28) to (6.32). The sea water salinity term is assumed constant. On the boundaries of the region the following conditions are given:

$$\text{at} \quad \Gamma: \quad \partial T/\partial n = 0, \quad \partial u_s/\partial n = 0, \quad u_n = 0, \tag{6.38}$$

$$z = H: \quad k_T\, \partial T/\partial N = 0, \quad k\, \partial u/\partial N = -\tau_\lambda^b, \quad k\, \partial v/\partial N = -\tau_\theta^b. \tag{6.39}$$

Here n and N are the directions of the normals to the shore outline and to the bottom surface, respectively, and τ_λ^b and τ_θ^b are the viscous stresses parametrized in the model by the relations

$$\tau_\lambda^b = R_H \int_0^H u\, dz, \quad \tau_\theta^b = R_H \int_0^H v\, dz. \tag{6.40}$$

The original system of equations is subjected to a number of transformations. First by a change of variables the region of integration is sent into a cylinder with a base Ω_0 and a unit height. Then, following Bryan [6.41] and Marchuk [6.1] a barotropic component of motion is separated out using formulas (6.3). As a result we arrive at a system of equations in averaged quantities, which is reduced to the vorticity equation, and at a system of equations with respect to deviations. The equations in the deviations are reduced, after being split into transport-diffusion and adaptation, to a sequence of problems to be solved by cyclic permutation as presented in Sec. 6.2. The problems can be written as

$$\begin{aligned} \partial u'/\partial t + \Lambda_u u' &= f_u, \\ \partial v'/\partial t + \Lambda_v v' &= f_v, \\ \partial T/\partial t + \Lambda_T T &= f_T \end{aligned} \tag{6.41}$$

at the transport-diffusion stage and as

$$\begin{aligned} \partial u'/\partial t + L_u u' &= g_u, \\ \partial v'/\partial t + L_v v' &= g_v, \\ \partial T/\partial t &= 0 \end{aligned} \tag{6.42}$$

at the adaptation stage. The structure of the operators Λ_u, Λ_v, and Λ_T in (6.41) coincides with that of the operators in (6.42). The right-hand sides of (6.41) contain some averaged advection and diffusion terms that arise in the equations owing to the separation into a barotropic and a baroclinic component of motion. The operators L_u and L_v in (6.42) contain terms with the Coriolis parameter. The right-hand sides g_u and g_v contain pressure gradients.

The method of solving the vorticity equation is based on combining the finite-element method and the splitting method. The procedure is described in [6.19], [6.20]. Equations (6.41) are reduced

to the form (6.6) by the method described in Sec. 6.2 and solved by the methods also described there.

To make numerical calculations according to the model a grid was constructed with coordinatewise condensations in latitude in the zones where the Gulf Stream and Kuroshia separate from the continents near the equator and in the area of the Antarctic Circumpolar Current. In longitude the condensation was effected near the western shores of the Atlantic and Pacific Oceans.

This choice was stipulated by the presence in the ocean of zones of narrow intensive currents which are very difficult to describe on a uniform grid with the aid of schemes which have no scheme viscosity since an oscillating computational mode appears in the solution. Condensing the grid may allow one to reduce the values of the turbulent exchange coefficients in the model and thus enhance the role played by the nonstationary factors and nonlinearity.

The grid spacing ranges from one degree to five degrees. Calculations for the barotropic part were carried out with the value of parameters $R = R_h = 0.5 \times 10^{-7}$ s^{-1} and $A_H = 5 \times 10^9$ cm^2/s. Calculations of the integral stream function with a variable bottom topography have shown that condensations of the grid allow one to reduce the values of the friction parameters in comparison with the calculations on a uniform five-degree grid [6.42]. Reducing the friction parameter leads to more intensive currents in the zones of boundary layers and to a greater space variability of the stream function field.

6.4.2. A model of the world ocean with surface turbulent layer. The mathematical statement of the problem is based on the solution of equations (6.28) to (6.32) without salinity. The nonlinear terms in the equations of motion are ignored. The density is given by the linear equation of state.

This model [6.43] basically differs from those in [6.42], [6.44] in that it takes into account the mixed layer whose dynamics is given by formula (6.27) and in the boundary conditions for the temperature at the surface (see (6.35)).

The numerical experiment has been carried out for a region of the world ocean with a five-degree resolution. Twenty eight levels were used in the vertical.

A numerical implementation uses a combination of explicit and implicit methods of approximating with respect to time, displaced grids [6.45] and the methods of space discretization suggested in [6.1].

At the points with integer indices i, j one finds the values of T, ψ, k and w (ψ being the stream function) and at the points with half-integer indices, the values of u and v. The value of the coefficient k at the points with half-integer indices is obtained by averaging k over four surrounding points. In the vertical T, u and v are calculated at the points with integer indices and w and k are calculated

at the points with half-integer indices. The boundary conditions are fulfilled at the points with half-integer indices.

The depth of the surface mixed layer h is determined at the first computational point l from the surface at which the condition

$$k = (cz_{l+1/2})^2 \sqrt{(\partial u/\partial z)^2 + (\partial v/\partial z)^2 - (g/\rho_0)\,\partial\rho/\partial z} \leqslant k_0$$

holds, after which k is calculated from formula (6.27) with $h = z_{l+1/2}$. An a priori value of k_0 is the coefficient of the vertical turbulent exchange below the mixed layer.

The method of solving the equations of motion assumes splitting the velocities into a barotropic component $\bar{u}$, $\bar{v}$ and a baroclinic component u', v':

$$(u,\ v) = (\bar{u},\ \bar{v}) + (u',\ v'),$$

$$(u',\ v') = \frac{1}{H}\int_0^H (u,\ v)\,dz = \left(-\frac{1}{a\rho_0 H}\frac{\partial\psi}{\partial\theta},\ \frac{m}{a\rho_0 H}\frac{\partial\psi}{\partial\lambda}\right).$$

The barotropic component is found from the solution of the equation in the stream function ψ. The value of the stream function in the Antarctic, which was appoximated by the latitudinal line, was found from the relation resulting from summation of the finite-difference equation over all the nodes of the grid, except that of the Antarctic.

In matrix form equations (6.28), (6.29) become

$$\frac{\partial \mathbf{U}}{\partial t} + G\mathbf{U} - \frac{\partial}{\partial z}k\frac{\partial \mathbf{U}}{\partial z} = \mathbf{\Phi}, \tag{6.43}$$

where

$$\mathbf{U} = \begin{pmatrix} u \\ v \end{pmatrix},\quad G = \begin{pmatrix} 0 & 2wn \\ -2wn & 0 \end{pmatrix},\quad \mathbf{\Phi} = \begin{pmatrix} -\dfrac{m}{a\rho_0}\dfrac{\partial p}{\partial\lambda} + F^{\lambda} \\ -\dfrac{1}{a\rho_0}\dfrac{\partial p}{\partial\theta} + F^{\theta} \end{pmatrix},$$

and the boundary conditions become

$$z = 0:\quad k\rho_0\,\partial\mathbf{U}/\partial z = -\boldsymbol{\tau};\quad z = H:\quad k\rho_0\,\partial\mathbf{U}/\partial z = -\boldsymbol{\tau}^b.$$

Approximating (6.43) in the vertical coordinate it is not hard to obtain the equation for the velocity shift $\mathbf{S}_k = \mathbf{U}_{k+1} - \mathbf{U}_k$. Approximating the equation obtained by an implicit scheme with respect to time we arrive at a system of algebraic equations that can be solved by matrix factoring.

Once the shift $\mathbf{S}$ and the stream function ψ have been found the velocities

$$\mathbf{U}_k = (1/H)\left[\sum_{l=1}^{K}\mathbf{U}_l\Delta z_l - \sum_{l=1}^{K-1}\mathbf{S}_l\,(H - z_{k+1/2})\right] + \sum_{l=1}^{k-1}\mathbf{S}_l$$

are calculated. The heat-transfer equation was approximated by an implicit scheme with respect to time. The divergent part has been approximated by an antisymmetric operator. The system of algebraic equations obtained is solved by the block Gauss-Seidel method. The vertical velocity is evaluated from the continuity equation. The hydrostatic equation is approximated as follows:

$$p_k = p_s + g \sum_{l=1}^{k} \rho_l \Delta z_l - (g/2)\, \rho_k \Delta z_k$$
$$p_b = p_s + g \sum_{l=1}^{K} \rho_l \Delta z_l \qquad (6.44)$$

where p_b is the bottom pressure and p_s is the free surface pressure. Equations (6.44) in the difference formulation correctly describe the kinetic-potential energy exchange.

Numerical results. The numerical experiment has been carried out on a Burrough 6700 computer for the water area of the world ocean between 67.5N and 67.5S. The ocean depth was assumed constant and equal to 3250 m. The calculated levels in the vertical practically coincide with the standard horizons. The outdoor observations for the four seasons were kindly provided by A. H. Oort and used as boundary values for the wind stress and ocean surface temperature. The data on a one-degree grid were averaged into nodes of five-degree squares. The interpolation between the seasons is linear. The computing program provides for the study of the role played by the bottom friction τ^b whose values have been found from the Ekman equations of motion.

The numerical experiment was conducted in two stages.

The calculation of the first stage began with the state of rest of the ocean and the climatic distribution of the temperature field. The boundary values at the ocean surface relate to the 15th of September (the summer season August-September-October according to [6.46]). This variant was calculated for eleven years of model time. At this stage there was an adaptation of hydrothermodynamic fields. In the process of computation the value of the coefficient A_H was reduced from 5×10^8 to 2.5×10^7 cm^2/s and that of k_0 from 10 to 1 cm^2/s. The coefficient of horizontal pulse exchange $A_H = 5 \times 10^9$ cm^2/s and the time step equalled 24 hours. The coefficient k in the formula for bottom friction was also changed from 10^2 to 1 cm^2/s.

The second stage consisted in solving the problem of seasonal variability with the initial data taken from the solution of the first stage. The nonstationary problem was calculated for 25 years of model time. For the last 20 years it is assumed that $\tau^b = 0$.

Figure 6.2 illustrates the temperature field distribution in the

horizontal plane. The upper field represents the initial state of the second stage, the lower field was obtained at the end of the second stage.

Along the western shores in the Northern Hemisphere there appeared an intensive condensation of isolines typical of jet streams. This

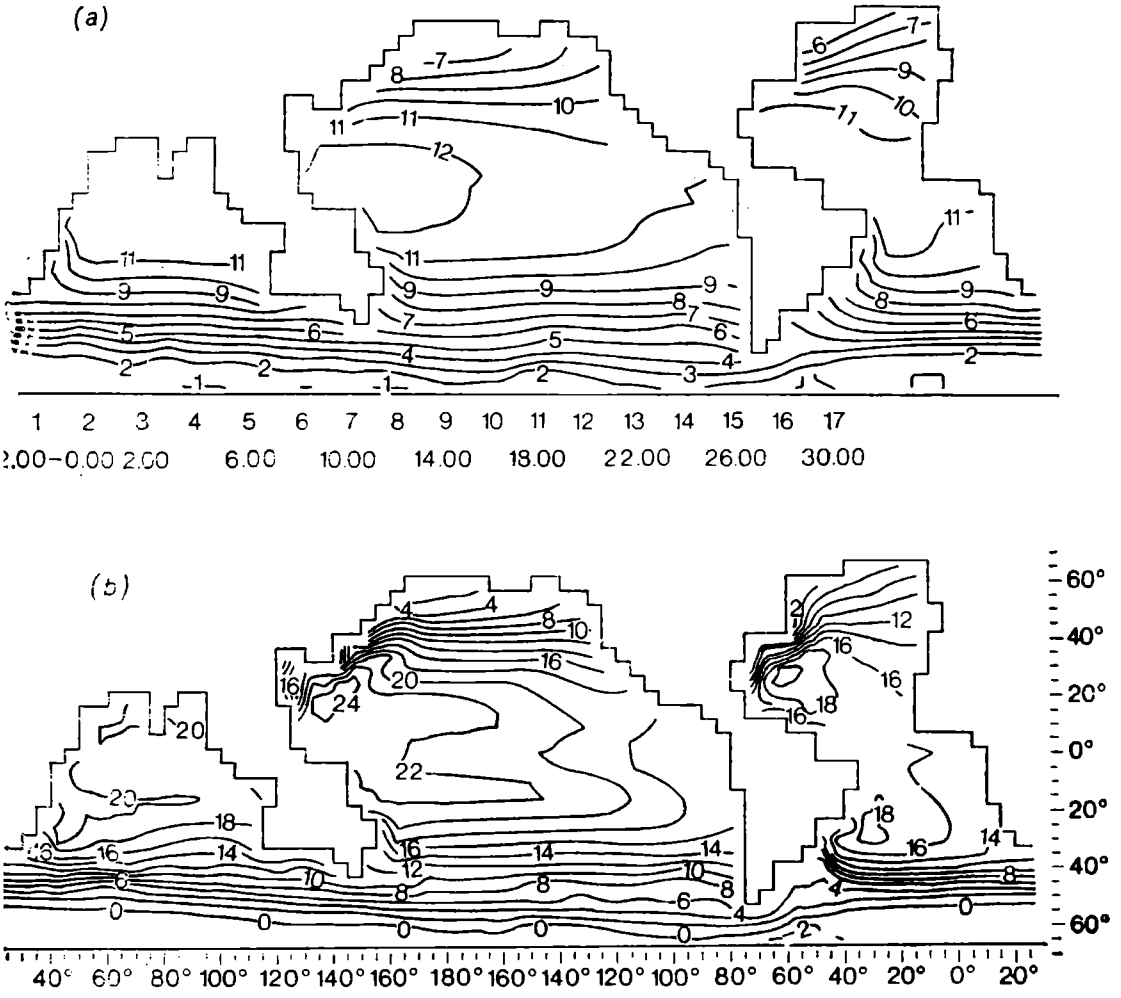


Fig. 6.2. Temperature field at a depth of 199.0 m, (*a*) the initial field and (*b*) the calculated field, September

justifies the conclusion that the trends resulting from the solution are not an interpolation in time of the initial state field. However, the short term of the time integration gives grounds to judge only about the trends in the solution obtained.

We discuss the results obtained from the problem of seasonal variability of the general circulation of the world ocean [6.21], [6.43], [6.47]. Figures 6.3 and 6.4 show the surface of the lower boundary of the UQL (it is evaluated from the depth at which the coefficient k assumes an a priori minimum value) for the entire world

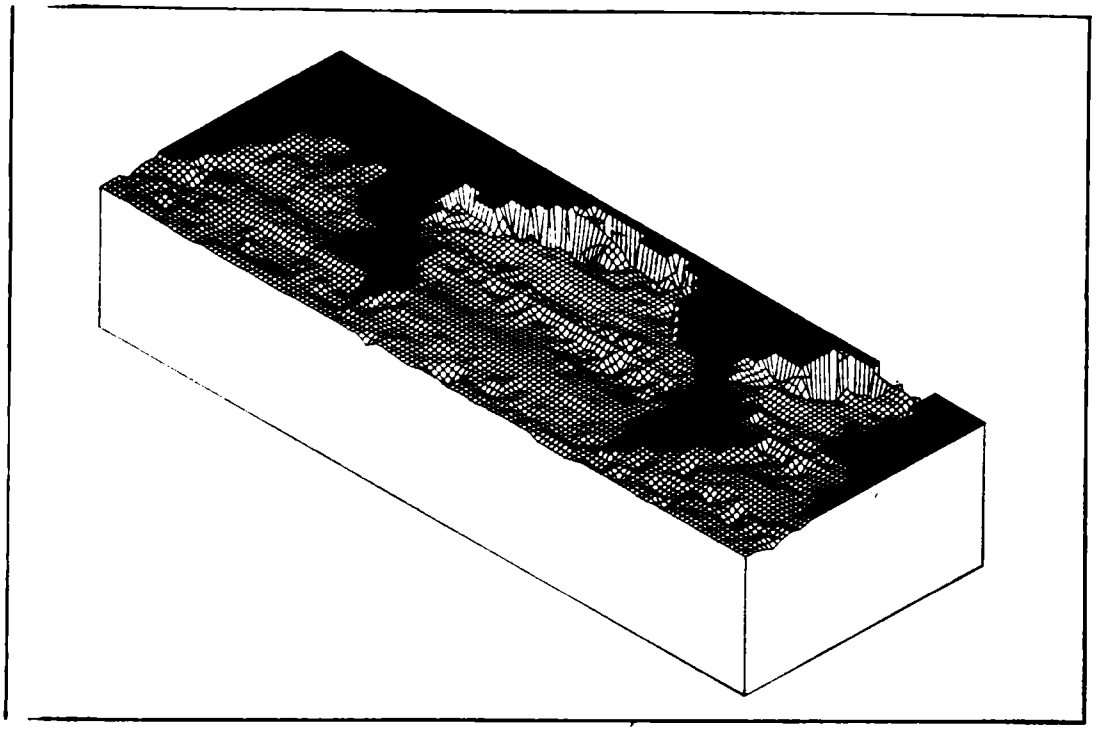


Fig. 6.3. Depth of the surface turbulent layer, January 15th; min — 8.5 m, max = 3000.0 m

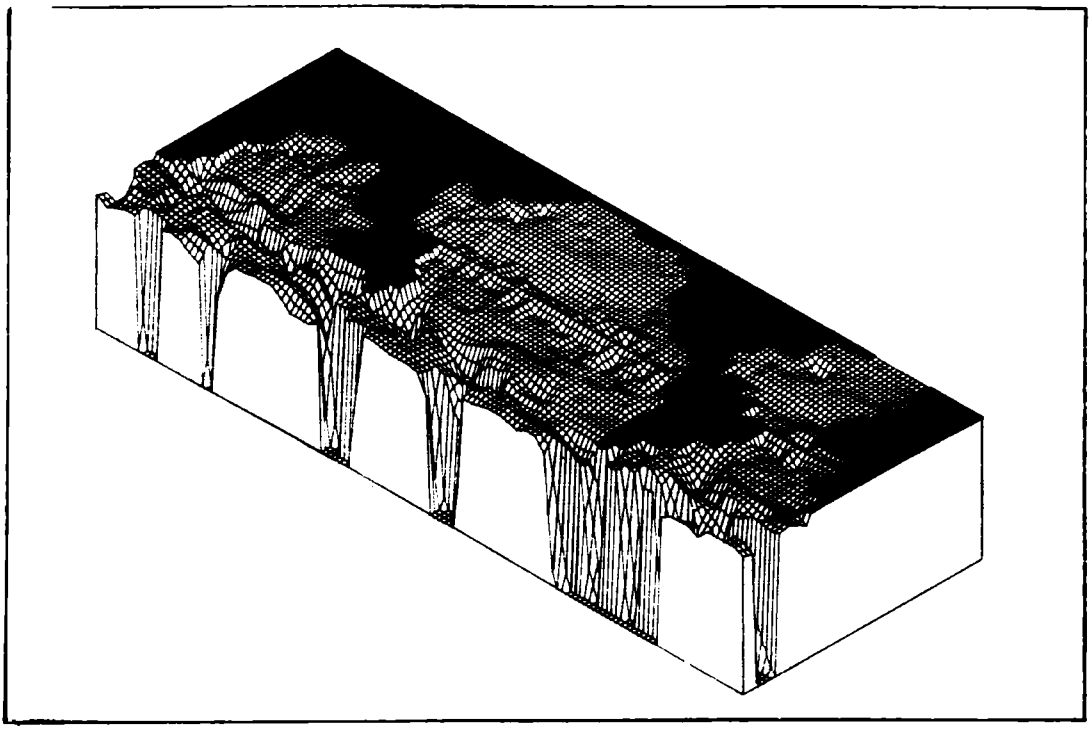


Fig. 6.4. The same as in Fig. 6.3, July 15th

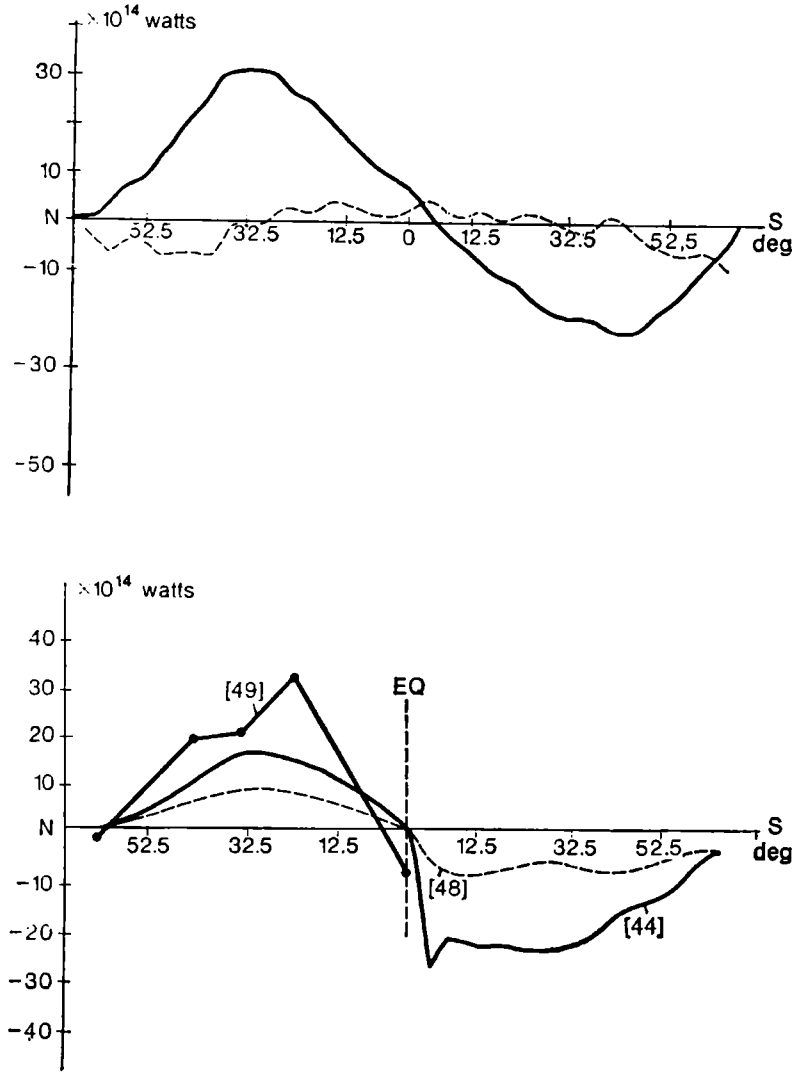


Fig. 6.5. Heat transport in the equator-pole direction (solid lines) and heat flow at the ocean surface (dashed lines). At the top: annual average distribution. At the bottom: calculated curves from [6.44], [6.48], and [6.49]

ocean, that was obtained in the seasonal trend problem. An analysis shows that in regions with a stable stratification the depth of the upper turbulent layer of the ocean is determined mainly by the local value of wind stress. To ascertain the role played by the vertical density gradient in these regions, it is necessary to improve the vertical resolution. In regions with a stable stratification the surface turbulent layer in the stationary problem does not exceed 70 m and in the seasonal trend problem it decreases owing to the determining influence of the increased vertical density gradient and is at most 40 m. In the regions where the turbulence penetrates from the surface down to greater depths it is of convective origin. Belonging to such regions are the high latitudes in the time of autumn-

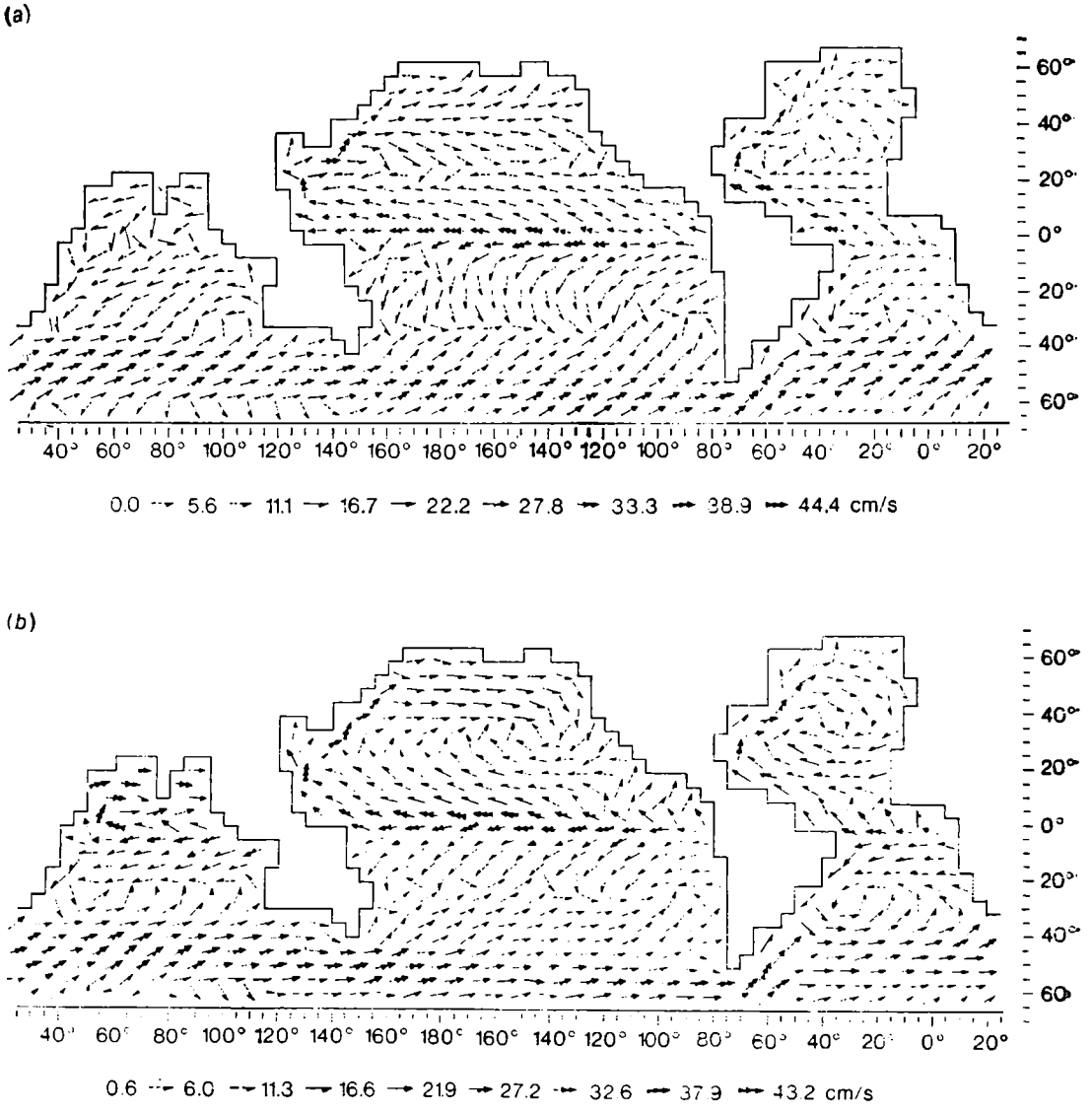


Fig. 6.6. Flow field at a depth of 2.5 m, (*a*) January, (*b*) July

winter cooling spell. It is interesting to note that the earliest beginning of convection and its latest termination were obtained in the areas of western boundary currents of the Northern Hemisphere. These are regions of the most intensive heat exchange with the atmosphere, the energy-active zones of the world ocean.

Of basic interest from the point of view of ocean-atmosphere interaction is heat transport by the ocean in the equator-pole direction

$$H_Q = c_p \rho_0 \int_0^{2\pi} \int_0^H (vT - (A_H/a)\,\partial T/\partial\theta)\,(a/m)\,d\lambda\,dz.$$

Figure 6.5 gives the calculated annual average distribution of this value for the entire world ocean. Also given for comparison are the calculated curves from [6.44], [6.48] and the estimate of heat transport obtained in [6.49] using the climatic data on the thermal balance of the atmosphere-underlying surface-ocean system. The character of the heat transport by the ocean agrees quite well with the existing representations. The quantitative deviations from the results of [6.44], [6.48] are related to the use of the turbulence model, good vertical resolution, and partly to the different boundary conditions at the surface for the temperature.

The field of surface flows is given in Fig. 6.6. All the year round intensive flows were obtained along the equator, in the ACC, and off the western shores. Significant seasonal changes in surface flows are observed. Note that at a depth of 200 m the flows along the equator change sign and at the depth of 3000 m the picture is opposite to the surface pattern.

6.4.3. The model of the formation of large-scale temperature and salinity fields in the world ocean. We now proceed to consider a model of the general circulation of the world ocean incorporating the calculation of temperature and salinity [6.50]. The equations of motion neglect the evolutionary and nonlinear terms and the horizontal turbulent exchange is parametrized in the form of the terms of mass friction with coefficient R. The equations of heat and salinity diffusion are used in their full form.

The boundary conditions at the lateral surface and at the bottom are chosen as follows:

$$\text{at} \quad \Gamma: \quad T = T^*, \quad S = S^*, \int_0^H u_n \, dz = 0; \tag{6.45}$$

$$z = H: \quad u = v = w = 0; \tag{6.46}$$

$$z = H_1: \quad T = T^*, \quad S = S^*. \tag{6.47}$$

Here $z = H_1 \leqslant H$ is the lower boundary of the baroclinic layer of the ocean below which the fluid is barotropic. The algorithm for solving this problem is as follows. At each time step two linearized equations for the temperature and salinity ($\varphi = T$ or S)

$$\frac{\varphi - \varphi^0}{\Delta t} + \frac{u^0}{\sin\theta}\frac{\partial\varphi}{\partial\lambda} + v^0\frac{\partial\varphi}{\partial\theta} + w^0\frac{\partial\varphi}{\partial z} = \frac{\partial}{\partial z}\left(k'\frac{\partial\varphi}{\partial z}\right) + A'_H\Delta\varphi \tag{6.48}$$

and an equation for the integral stream function ψ

$$-\varepsilon'\Delta\psi + A(\lambda, \theta)\frac{\partial\psi}{\partial\lambda} + B(\lambda, \theta)\frac{\partial\psi}{\partial\theta} = F(\lambda, \theta) \tag{6.49}$$

are solved. In (6.48) u^0, v^0, w^0, and φ^0 are referred to a previous instant of time t, equation (6.49) is obtained from equations (6.28) to (6.31) and the corresponding boundary conditions. Once T, S, ψ, and ρ have been evaluated from (6.48), (6.49) and (6.34), the velocity vector components u, v and w are found from the simple formulas [6.51], [6.52] at time $t + \Delta t$ and then the next time step is taken. Small parameters have appeared in equations (6.48) and (6.49) written in dimensionless variables: ε' characterizes the ratio of the Ekman layer thickness to the baroclinic layer depth H_1 and k' and A'_H represent the ratio of the value of the diffusive transport to that of the advective one.

Equations (6.48), (6.49) are solved according to the scheme of a second order of approximation (6.23).

The region of the world ocean was approximated on a uniform five-degree grid displaced relative to the equator by 2.5°. In the vertical a nonuniform grid 0, 100, 363, 773, 1320, and 2000 m has been chosen which became uniform when a new variable $\eta = z^p$ was introduced in the vertical. Solutions of stationary diffusion equations were given as the initial fields T^* and S^*. The following values of the physical parameters were used: $k = 75$, $R = 1.5 \times 10^{-7}$, $k_T = 1$, $A_H = 3.6 \times 10^7$ (in CGS). The coefficient of the horizontal turbulent diffusion in longitude near the equator in the Pacific and Atlantic Oceans was increased to 3.6×10^9 cm^2/s. Thus the insufficient advective transport of heat and salt by the equatorial currents, which turn out to be weak due to the crude space resolution, was replaced by the diffusive transport. The value of the coefficient R near the equator in the Indian Ocean was increased to 0.73×10^{-4}. The computation was carried out with the time step of 30 days. The problem has been calculated for thirteen years. During that time the mean-square value of the density anomaly, obtained by averaging over all computational grid nodes, was monotonic decreasing. The relative increment of the value at the end of the computation was four orders smaller than at the beginning.

Figure 6.7 represents the vertical structure of the calculated temperature and salinity fields in the meridional plane at the initial instant of time (the dashed curves) and in thirteen years of model time. The displacements of subtropical temperature maxima with increased depth toward the high latitudes are clearly seen. There is a qualitatively realistic thermocline whose thickness significantly decreases near the equator. This is one of the features of the general ocean circulation in subtropical turnovers. Qualitatively truly simulated are the salinity fields with a satisfactory halocline. There is a weak intrusion of Antarctic waters of decreased salinity in both fields of Fig. 7*b*.

Figure 6.8 presents the horizontal distribution of the calculated temperature and salinity at the level of 363 m. There are five local

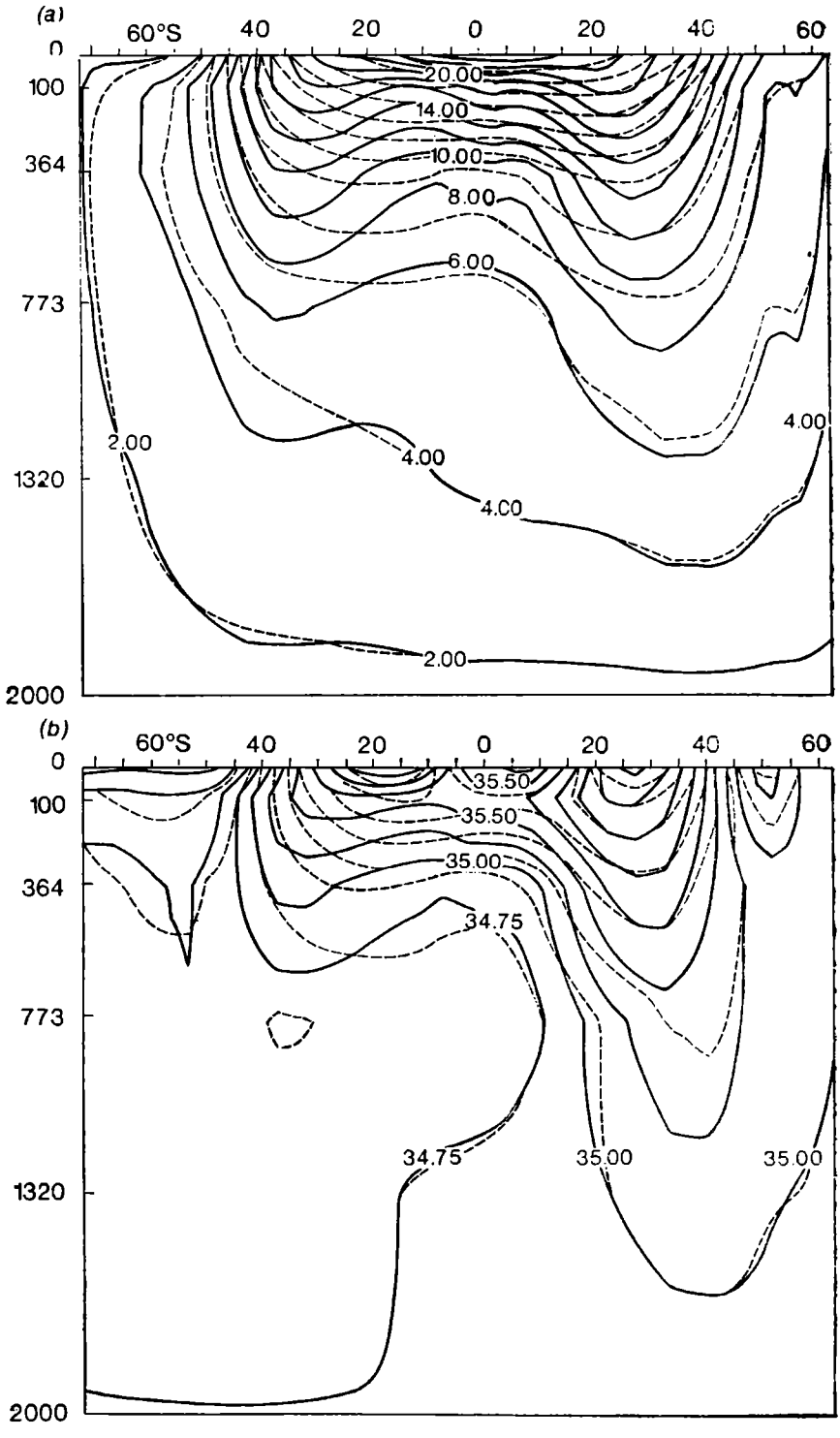


Fig. 6.7. Zone-averaged temperature fields (*a*) and salinity fields (*b*) in the Atlantic Ocean at the zero time (dashed lines) and in thirteen years of model time (solid lines)

subtropical temperature maxima in the northern and southern parts of the oceans. In the salinity fields, as in the calculated temperature

fields, qualitatively truly simulated are the zones of subtropical large-scale circulation systems with increased salinity.

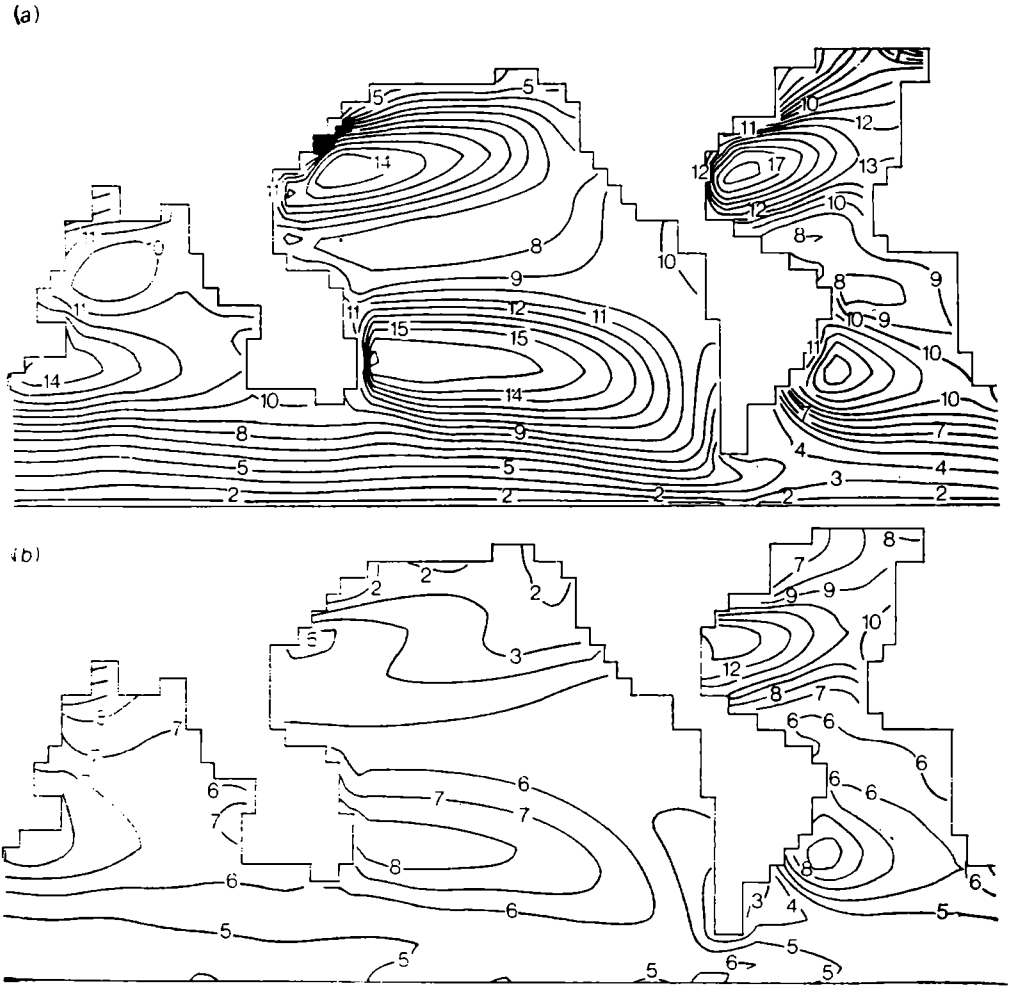


Fig. 6.8. Calculated temperature fields (*a*) and salinity fields (*b*) at a depth of 363 m. The figures on the isotherms refer to Celsius degrees, the figures on the isohalines refer to their numbers; the salinity per mil is $34.2 + 0.2\,(n - 1)$, where n is the number of the corresponding isoline

6.5. DIAGNOSTIC AND SEMIDIAGNOSTIC CALCULATIONS OF HYDROLOGICAL CHARACTERISTICS OF THE WORLD OCEAN

We have discussed above some numerical methods for simulating the characteristics of the ocean. They will develop and will lead in the future to safer conclusions about the dynamics of the ocean, but at present there are objective difficulties which somewhat limit the possibilities of the above models. Complicated, for example, is the problem of parametrizing the turbulent diffusion

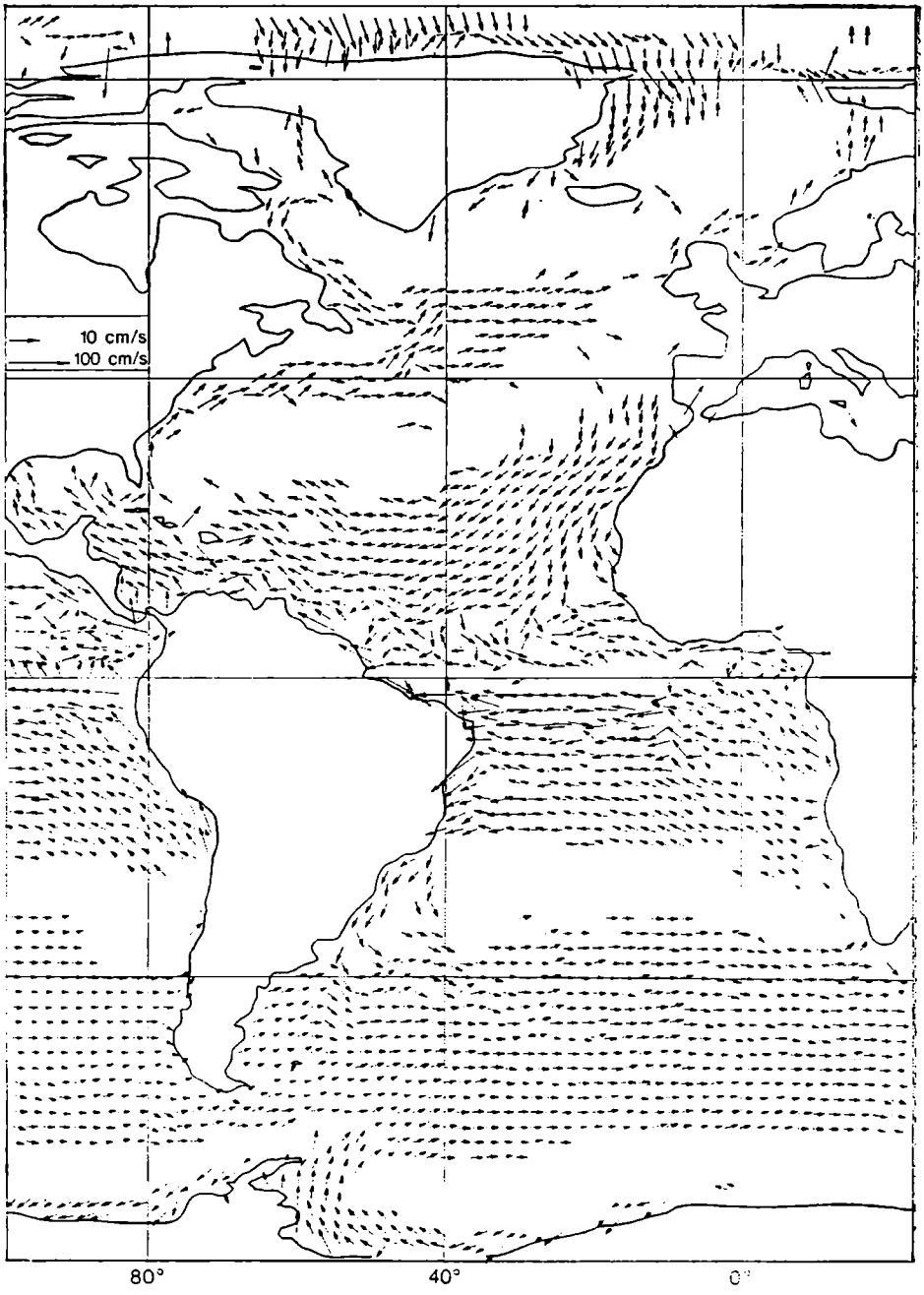


Fig. 6.9. Velocity of surface flow in the Atlantic Ocean in the summer season. The arrows are given only for the areas where flow velocity is at least 10 cm/s

of heat and salt in the ocean. These processes are of great importance, but at present it is still difficult to take them accurately into account. The presence of vortex motions of meandering jet currents adds to the difficulties. Along with the directions indicated above it is appropriate, therefore, to develop also another trend in the studies associated with the use of observations. The numerical models and cal-

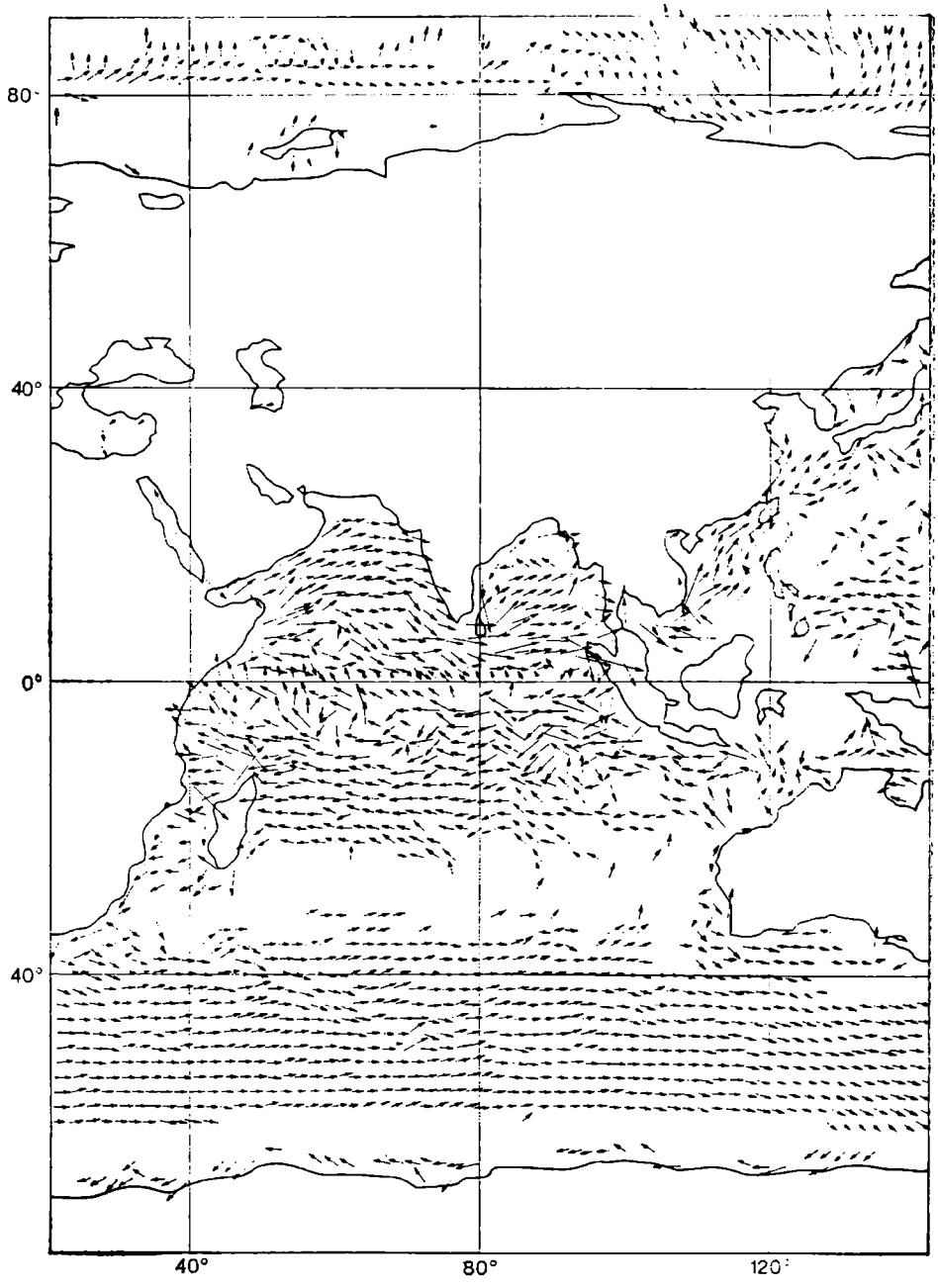


Fig. 6.10. Velocity of surface flow in the Indian Ocean in the summer season. The scale of flow velocity is the same as in Fig. 6.9

culations of currents of the world ocean and individual basins given observed temperature and salinity fields were published earlier [6.53], [6.54]. By way of illustration Figs. 6.9 and 6.10 give surface flow charts for a summer season in the Atlantic and Indian Oceans. The calculations were carried out for 31 horizons of the world ocean (from surface to bottom) using a one-degree spacing in the horizontal.

Of course, the currents obtained (up to 1 m/s) are 3 to 5 times more intensive than when completely simulated with a large spacing in the horizontal. In addition there are no constraints on diagnostic calculations on the part of modern computers. The density field is in general not adapted to wind fields and bottom topography and this is the main drawback of diagnostic calculations.

The next step embraces semidiagnostic calculations done in two stages. The first stage is purely diagnostic and the second is that of adaptation. At the latter the equations of the diagnostic model are supplemented with the equations of heat and salt transport (or the density equation) with the corresponding boundary conditions. The solution of the diagnostic problem is taken to be the initial estimate and a hydrodynamic adaptation of the temperature and salinity (or density) field and flow velocity to the wind field and bottom topography is carried out. The most acceptable method for the temperature and salinity (or density), it appears, is to solve mixed boundary-value problems: with the given temperature and salinity (or density) distribution at the ocean surface and open boundaries and with zero fluxes of properties, at the shore boundaries and bottom. The information about the distribution of those properties at the surface is the most extensive and rapidly growing, particularly that for the temperature. The fluxes of properties at the surface are known much more poorly. As the criterion of adaptation one assumes a criterion of the type

$$\mathscr{E}(t) = \frac{\partial F/\partial t}{\sqrt{(\partial F/\partial x)^2 + (\partial F/\partial y)^2}} = \min \quad \text{or} \quad \mathscr{E}(t) = \mathscr{E}_0 \qquad (6.50)$$

(where F is the temperature, salinity, or density) rather than the criterion of reaching the steady state in the ordinary mathematical sense (the behaviour with time of maximum discrepancies or kinetic energy). This criterion is based on the assumption that at the first stage of reaching the steady state the numerator, $\partial F/\partial t$, decreases much quicker than the denominator (the horizontal gradient of a property). At this stage the latter may decrease very insignificantly or even grow (depending on the degree of smoothness and the adaptation of the original fields). After a sufficiently long-time integration the second stage comes at which the rate (the rapidity) of the numerator decrease gradually diminishes and becomes lower than that of the denominator decrease. As a result the time trend of the curve $\mathscr{E}(t)$ (or the envelope of that curve if there are wave disturbances) gradually changes in such a way that either a minimum of $\mathscr{E}(t)$ is reached or some interval of the sharpest change $d\mathscr{E}/dt$ appears in the process of adaptation. Finally, a case is also possible with weakly expressed changes in the trend of the curve $\mathscr{E}(t)$ where the process stops simply upon reaching a sufficiently small value $\mathscr{E}_0$.

Physically, the implementation of any of the three cases means

that some optimum state has been reached where the fields are in the main already adapted but the property gradients are not yet very strongly smoothed by the dissipative factors (physical and computational). Experience with calculations shows that with a very long integration we obtain more and more strongly smoothed weak-gradient fields. To a considerable extent this is due, of course, to the role played by the "viscous terms" whose influence manifests itself much more slowly than that of the advective factors. Here we give only one of the possible forms of the adaptation criterion of the suggested type. The most convenient correct form of the criterion, which most clearly shows the transition from the first stage of adaptation (with a quick adaptation of the fields and a slow smoothing of gradients) to the second stage (with a relatively slow adaptation of the fields and a relatively quicker smoothing of the gradients), can be found, of course, only on the basis of numerical experiments. The possible variations may relate either to the character of averaging the function $\partial F/\partial t$ or to a particular form of the denominator of expression (6.50). One may use the horizontal velocity gradient instead of the horizontal density gradient or a combination of both in the denominator. It is possible that the averaging should be more conveniently carried out over each horizon separately, rather than over the bulk. In that case the transition section in each horizon would correspond to a different time of integration and would therefore be more noticeable. It would be possible to stop the process depending on what layers we want to adapt first of all. There are also prerequisites for more special forms of averaging; for example, averaging over an ensemble of significant gradients, i.e. gradients that are at least equal to a chosen value. There are also other variants whose principle, however, remains invariable. The appropriateness of such an approach is also suggested by the results of the semidiagnostic calculations (diagnosis plus adaptation) we have lately carried out. It has turned out that if in solving the adaptation problem one takes the results of the diagnostic calculation as the initial estimate, then the process of the adaptation of the fields develops for some time (lasting from some days to some months depending on the scale of the grid and the hydrology of the basin) in such a way that the strong well-defined jet streams become relatively even more intensive and the wide slow background streams become even weaker. In two cases out of the three (the North Atlantic and the Black Sea) the intensification of the jets was only relative, at the expense of the much more pronounced weakening of the background circulation (absolute velocities in the jet streams have even somewhat decreased) and in the third case (the Lake of Sevan) an absolute increase in the velocities occurred. It is obvious that with a sufficiently long integration the dissipative factors will result in an ever increasing smoothing of the solution. As an example Fig. 6.11 gives a chart of the

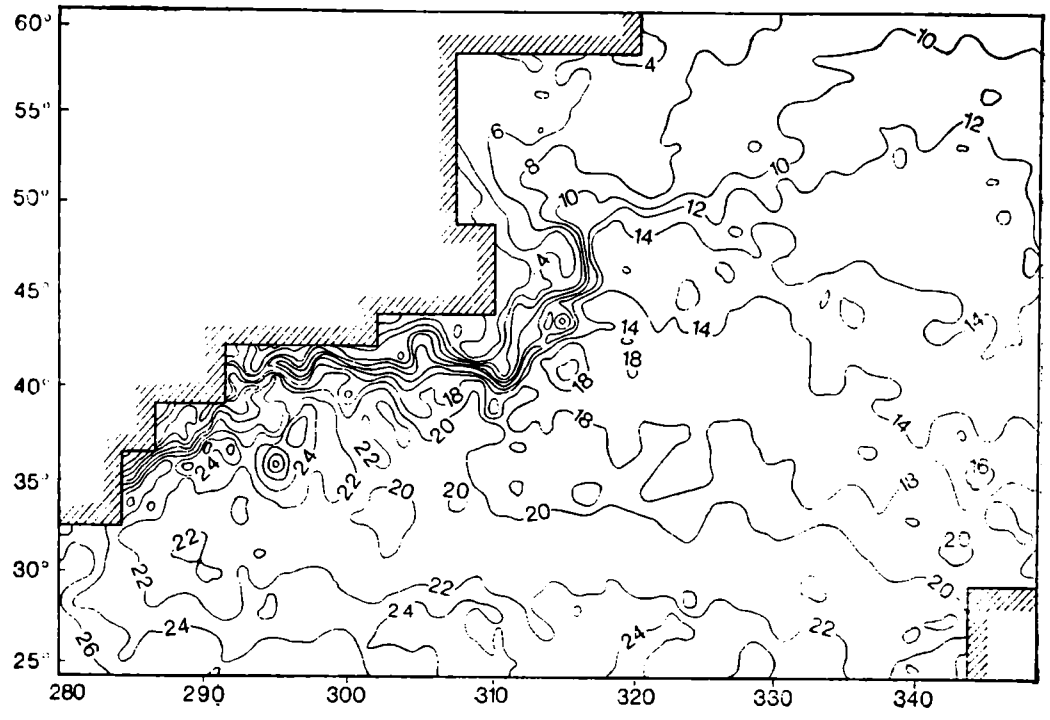


Fig. 6.11. Climatic temperature field at a level of 50 m (summer)

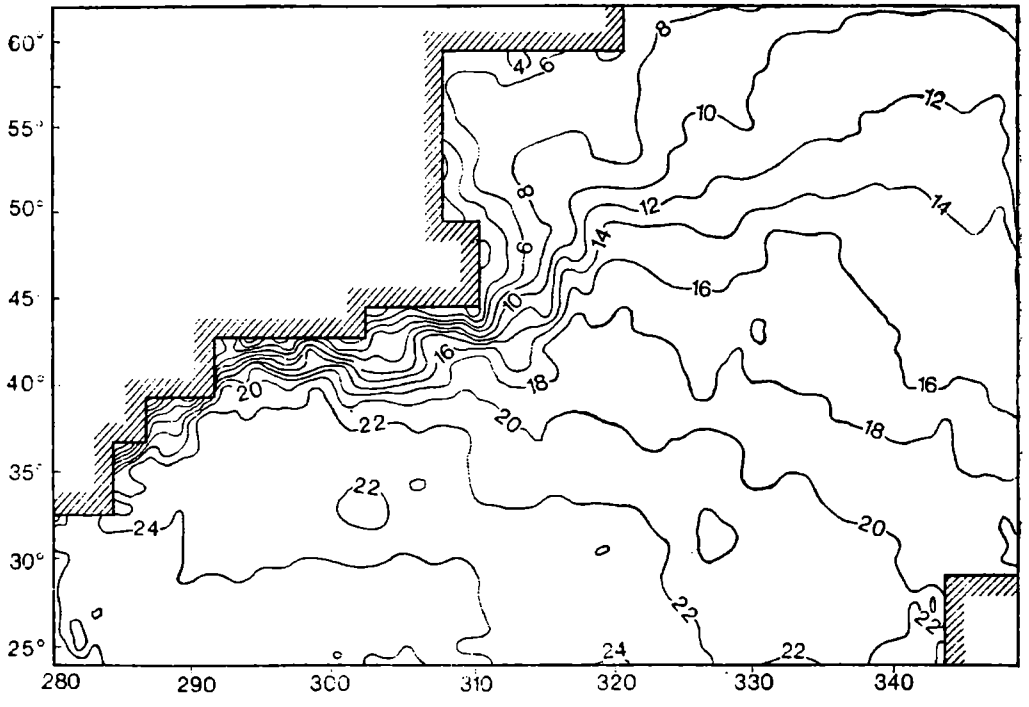


Fig. 6.12. Temperature field at a level of 50 m. The result of semidiagnostic adaptation (65 days of model time)

original climatic temperature field of the North Atlantic at the level of $z = 50$ m. Besides useful information the field contains some noise, the errors and asynchronism of measurements, the effect of small nonstationary eddies, etc. The chart of Fig. 6.12 is adapted to the bottom topography, the wind field, and the equation and is therefore better ordered, although smoothed, a spacing of 1° is too large for the areas of jet streams. The temperature, salinity, and flow fields obtained by semidiagnostic calculations will serve as a basis for the analysis of cause-effect relations and as initial fields for a four-dimensional analysis and prediction of sea currents.

References

6.1. G. I. Marchuk, *Numerical Solution of Problems of the Dynamics of the Ocean and Atmosphere* (Leningrad: Gidrometeoizdat, 1974) (in Russian).

6.2. G. I. Marchuk and B. A. Kagan, *Oceanic Tides* (Leningrad: Gidrometeoizdat, 1977) (in Russian).

6.3. L. V. Ovsyannikov, *Uniqueness Theorem for a Linearized System of Ocean Dynamics Equations* (Novosibirsk: MGU, Preprint, 1966) (in Russian).

6.4. A. A. Kordzadze, *Dokl. Akad. Nauk SSSR* 219, No. 4: 856-859 (1974) (in Russian).

6.5. M. A. Bubnov and A. V. Kazhikhov, *Dokl. Akad. Nauk SSSR* 198, No. 4: 801-804 (1971) (in Russian).

6.6. V. P. Kochergin, in: *Numerical Methods of Calculating Oceanic Tides* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1974): 43-52 (in Russian).

6.7. V. P. Kochergin, *Oceanic Tides: Theory and Methods of Calculation* (Moscow: Nauka, 1978) (in Russian).

6.8. A. A. Kordzadze, *Dokl. Akad. Nauk SSSR* 244, No. 1: 40-43 (1979) (in Russian).

6.9. A. A. Kordzadze, *Mathematical Modelling of the Dynamics of the Ocean* (*Mathematical Modelling of the Dynamics of the Ocean and Atmosphere*), Part One (Novosibirsk: Nauka, 1980): 114-139 (in Russian).

6.10. A. A. Kordzadze, *Mathematical Questions of Solving Ocean Dynamics Problems* (Novosibirsk: Vychisl. Tsentr. SO AN SSSR, 1982) (in Russian).

6.11. G. I. Marchuk and M. A. Bubnov, *Dokl. Akad. Nauk SSSR* 248, No. 2: 836-837 (in Russian).

6.12. A. A. Kordzadze, *Dokl. Akad. Nauk SSSR* 232, No. 2: 308-311 (1977) (in Russian).

6.13. A. A. Kordzadze, *Dokl. Akad. Nauk SSSR* 237, No. 1: 52-55 (1977) (in Russian).

6.14. M. A. Bubnov, *On the Solvability of Tide Dynamics Problems in a Stratified Ocean* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint No. 70, 1977) (in Russian).

6.15. M. A. Bubnov, in: *Mathematical Modelling of the Dynamics of the Ocean* (*Mathematical Modelling of the Dynamics of the Ocean and Atmosphere*), Part One (Novosibirsk: Nauka, 1980): 59-75 (in Russian).

6.16. G. I. Marchuk and M. A. Bubnov. *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 16, No. 3: 211-218 (1980) (in Russian).

6.17. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd enl. ed. (Moscow: Nauka, 1980) (in Russian).

6.18. G. I. Marchuk and V. I. Kuzin, *The Scheme of Solving a Two-dimensional Parabolic Equation by the Finite-element Method in Combination with the Splitting Method* (Moscow: Otdel Vychisl. Matem. AN SSSR, Preprint, 1982) (in Russian).

6.19. V. I. Kuzin, in: *Mathematical Modelling of the Dynamics of the Ocean* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1982): 156-169 (in Russian).

6.20. V. I. Kuzin, in: *Numerical Solution of Ocean Dynamics Problems* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1982): 45-52 (in Russian).
6.21. G. I. Marchuk *et al.*, *Mathematical Models of Circulation in the Ocean* (Novosibirsk: Nauka, 1980) (in Russian).
6.22. G. I. Marchuk, A. A. Kordzadze, and Yu. N. Skiba, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 11, No. 4: 379-393 (1975) (in Russian).
6.23. G. I. Marchuk, M. A. Bubnov, V. B. Zalesny, and A. A. Kordzadze, in: *Problems of Computational Mathematics* (Novosibirsk: Nauka, 1983) (in Russian).
6.24. V. V. Shaidurov, *Chisl. Metody Mekh. Sploshn. Sredy* 3, No. 2: 96-104 (1972) (in Russian).
6.25. E. A. Volkov, *Diff. Uravneniya* 1, No. 8: 1070-1084 (1965) (in Russian).
6.26. G. I. Marchuk and V. V. Shaidurov, *Increasing the Accuracy of Solutions of Difference Schemes* (Moscow: Nauka, 1979) (in Russian).
6.27. V. P. Kochergin and A. V. Shcherbakov, *Chisl. Metody Mekh. Sploshn. Sredy* 5, No. 1: 88-97 (1974) (in Russian).
6.28. A. V. Shcherbakov and V. P. Kochergin, *Chisl. Metody Mekh. Sploshn. Sredy* 8, No. 2: 125-129 (1977) (in Russian).
6.29. V. P. Kochergin, V. I. Klimok, and A. V. Shcherbakov, *Chisl. Metody Mekh. Sploshn. Sredy* 8, No. 3: 105-111 (1977) (in Russian).
6.30. G. I. Marchuk, V. P. Kochergin, V. I. Klimok, and V. A. Sukhorukov, *J. Phys. Oceanogr.* 7, No. 7: 865-875 (1977).
6.31. A. N. Kolmogorov, *Dokl. Akad. Nauk SSSR* 6, Nos. 1-2: 56-58 (1942) (in Russian).
6.32. A. I. Zagustin, *Trudy Voronezh. Univ.* 10, No. 2: 7-39 (1938) (in Russian).
6.33. P. Y. Chou, *Quart. Appl. Math.* 3, No. 1: 38-54 (1945).
6.34. B. I. Davydov, *Zh. Eksp. i Teor. Fiz.* 35: 527-529 (1958) (in Russian).
6.35. F. H. Harlow and P. J. Nakayama, *Transport of Turbulence Energy Decay Rate*, Los Alamos Scientific Laboratory of the University of California, 1968, LA-3854, pp. 1-7.
6.36. K. Hanjalic and B. E. Launder, *J. Fluid Mech.* 52: 609-634 (1972).
6.37. V. P. Kochergin and V. A. Sukhorukov, *Chisl. Metody Mekh. Sploshn. Sredy* 6, No. 5: 81-93 (1975) (in Russian).
6.38. A. M. Obukhov, *Trudy Inst. Teor. Geofiz. Akad. Nauk SSSR* 24 (151): 3-42 (1946) (in Russian).
6.39. V. P. Kochergin, V. I. Klimok, and V. A. Sukhorukov, *Chisl. Metody Mekh. Sploshn. Sredy* 8, No. 5: 102-114 (1977) (in Russian).
6.40. G. I. Marchuk, V. P. Kochergin, V. I. Klimok, and V. A. Sukhorukov, *Dokl. Akad. Nauk SSSR* 247, No. 1: 68-72 (1979) (in Russian).
6.41. K. Bryan, *Dynamics of Atmospheres and Oceans*, No. 3: 327-338 (1979).
6.42. V. I. Kuzin, *Chisl. Metody Mekh. Sploshn. Sredy* 5, No. 4: 71-84 (1974) (in Russian).
6.43. J. Siindermann, V. P. Kochergin, V. I. Klimok, V. A. Sukhorukov, and G. Fridrikh, *Mathematical Modelling of the Seasonal Variability of the General Circulation of the World Ocean Taking into Account the Surface Turbulent Layer* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint No. 381, 1982) (in Russian).
6.44. K. Takano, Y. Mintz, and Y. J. Han, *2nd Conference of Prediction, AMS, Oct. 1973, Monterey, California* (unpublished).
6.45. K. Bryan, *J. Comput. Phys.*, No. 3: 347-376 (1969).
6.46. S. Levitus and A. H. Oort, *Bull. Amer. Meteor. Soc.* 58, No. 12: 1270-1284 (1977).
6.47. G. I. Marchuk, V. P. Kochergin, V. I. Klimok, and V. A. Sukhorukov, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 12, No. 8: 841-849 (1976) (in Russian).
6.48. K. Bryan and L. Lewis, *A Water Mass Model of the World Ocean*, GFDL (NOAA), Princeton Univ. 1979: 1-81.

6.49. A. H. Oort and Vonder Haar T. H., *J. Phys. Oceanogr.* 6, No. 6: 781-800 (1976).
6.50. V. P. Kochergin and A. V. Shcherbakov, *Numerical Simulation of Temperature and Salinity Fields of the World Ocean* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint No. 398, 1982) (in Russian).
6.51. A. S. Sarkisyan, *Numerical Analysis and Prediction of Sea Currents* (Leningrad: Gidrometeoizdat, 1977) (in Russian).
6.52. A. S. Sarkisyan, *The Diagnostic Calculations of a Large-Scale Oceanic Circulation. The Sea.* Vol. 6 (New York: Wiley, 1977): 363-458.

7 Optimization Models for Environmental Protection

V. V. Penenko, A. E. Aloyan,
N. N. Obraztsov, A. V. Panarin,
A. V. Protasov, and V. F. Raputa

The paper presents methods for constructing numerical models in the environmental protection. The methods can be used for optimal planning of large-scale experiments, for the reconstructing fields from observed data, and for optimizing the quality of the atmosphere. The computing algorithms are constructed using the variational principle in conjunction with the splitting method. The structure of the models is determined by the systems of basic and adjoint equations of pollutant transport in the atmosphere, by the methods of calculating the functions of sensitivity, by the set of the functionals to be optimized, and by the collection of criteria for the control of the quality of the atmosphere.

7.1. SIMULATION OF POLLUTANT TRANSPORT IN THE ATMOSPHERE

Man-environment interaction and the assessment of its possible consequences have recently attracted a great deal of scientists' attention in different countries. The intensification of technological progress, difficulties in carrying out full-scale studies of actual processes that occur under the joint action of natural and human factors, their high cost, and the irreversibility of many experiments in ecology and socio-economic areas are actively stimulating the development of the basic theory and methods of mathematical modelling for tackling questions associated with the environmental protection problem.

This paper gives a brief account of the investigations carried out at the Computing Centre of the Siberian Division of the USSR Academy of Sciences and in the Department of Computational Mathematics of the USSR Academy of Sciences under the guidance of Academician G. I. Marchuk into the methods of mathematical modelling in the environmental protection problem. The principles of constructing mathematical models for solving this class of problems are formulated in [7.1]. In this paper attention is focused on developing these principles for problems that arise in working out systems of optimal control of the quality of the natural surroundings [7.2], [7.3] using models for the study of pollutant transport in the atmosphere as an example. In particular, methods for estimating

the effects of different factors on the atmosphere pollution and the state of the atmosphere from experimental data, for designing observational experiments, and for monitoring sources of pollution of the atmosphere are considered.

By structure the methods and models presented in this paper form a constituent part of the complex of models oriented at solving problems of the physics of the atmosphere and ocean and environmental protection and implemented as a general-purpose software package.

Mathematical models of this class are based on the equations of the hydrothermodynamics of the atmosphere and water objects taking simultaneously into account natural and human factors, on the equations of the transport and transformation of pollutants in the air and in the water, and on the equations and relations defining direct and feedback couplings in the system and conditions of the interaction of the atmosphere and pollutants with the nonuniform Earth's surface. These equations provide in differential form the laws of the conservation of energy, mass, and momentum, which together with the laws of thermodynamics and chemistry describe the processes occurring in the atmosphere, ocean, and soil and their interaction, provided their source is solar energy and the sources and sinks of heat and anthropogenic pollutants are known.

Essentially a complex of mathematical models, consistent with respect to the space-time scales, for the description of the processes under study is considered. The constructive implementation of the complex is effected using the splitting method as a basis [7.1]-[7.8]. The fundamental property of the splitting method is that different elements of mathematical models may be given a formally independent consideration at individual stages of splitting, but the structure of the entire complex must be such that at all the stages of numerical simulation the models preserve their primary physical meaning. This can be achieved by a simultaneous use of the variational principles and the splitting method in constructing numerical models. This combination makes apparent the best qualities of both types: the splitting method simplifies the construction and implementation of a model and the variational principle ensures the consistency of all the stages of numerical simulation [7.8].

For definiteness and brevity we consider a fragment of the complex of models, specifically the model of pollutant transport in the atmosphere, and demonstrate the formulations of primal problems and the structure of computing alogorithms for solving them. In the general structure of the complex it may be interpreted as one of the splitting stages.

Consider a three-dimensional region D on a spherical earth and introduce on it a Cartesian system of x-, y-, z-coordinates. Assuming the elements of hydrometeorological conditions to be given we write

the equations for pollutant transport as follows [7.1], [7.3], [7.9]:

$$L\varphi \equiv \frac{\partial \varphi}{\partial t} + \mathbf{u}\,\mathrm{grad}\,\varphi + A(\varphi) - \frac{\partial}{\partial z}\nu\frac{\partial \varphi}{\partial z} - \mathrm{div}_S \mu\,\mathrm{grad}_S \varphi = f(\mathbf{x}, t) \quad (7.1)$$

with conditions

$$l\varphi \equiv \nu\,\partial\varphi/\partial z - \beta\varphi = g_S \quad \text{for} \quad z = z_S(x, y) \quad (7.2)$$

$$\nu\,\partial\varphi/\partial z = -k_1(\varphi - \bar{\varphi}) \quad \text{for} \quad z = z_H, \quad (7.3)$$

$$\mu\partial\varphi/\partial n = k_2(\varphi - \bar{\varphi}) \quad \text{for} \quad (x, y, z) \in \sigma, \quad (7.4)$$

$$\varphi\,|_{t=0} = \varphi^0(\mathbf{x}).$$

Here $\mathbf{x} = (x, y, z)$ is a point of the region $D = \{S \times [z_S \leqslant z \leqslant z_H]\}$, S the range of horizontal (x, y) coordinates; n the direction of the normal to the lateral surface σ of D; z_S the surface topography of the Earth; z_H the upper boundary of the air mass, $\mathbf{u} = (u, v, w)$ the velocity vector of air particles; ν and μ the coefficients of vertical and horizontal turbulence; $f(\mathbf{x}, t)$ the distribution function of pollutant sources, g_S the distribution of pollutant surface sources, $\bar{\varphi}$ the value of background pollutant concentration at the boundaries of D; β, k_1, and k_2 the functions defining the conditions of the interaction of pollutants with the Earth's surface and with background processes; and A the operator describing local processes of the pollutant transformation. The subscript S is used to distinguish the operators acting only in the horizontal directions x and y. In particular, the range S may coincide with the entire Earth's surface. In that case the condition of periodicity of all functions on the surface of a sphere is specified instead of condition (7.4) on the lateral surface of D and spherical coordinates are used instead of Cartesian (x, y) coordinates.

The hydrometeorological conditions in D are taken into account via the velocity field $\mathbf{u}$ and via the turbulence coefficients. We consider them to be the input parameters of model (7.1) to (7.4). The rate of gravitational settling or rise of particles will be taken into account in the vertical component of the vector $\mathbf{u}$.

Depending on the purpose of investigation varied methods of specifying hydrometeorological conditions are used:

calculation from actually observed data;

simulation via models of the hydrothermodynamics of the atmosphere-ocean-soil system;

calculation via models in the regime of acquiring observed data;

simulation using distribution functions calculated from actual data.

The part of the function of state in the model in question is played by the distribution function of pollutant concentrations φ. In the actual atmosphere the pollutants form a multicomponent medium

and in order to describe the processes of their transport using models of the type (7.1) to (7.4) one should employ vector-matrix notation. The dimensions of vectors and matrices are determined by the number of pollutant different components. For example, if there are n substances, then instead of the function φ one introduces a vector function $\boldsymbol{\varphi} = \{\varphi_i, (i = \overline{1, n})\}$, where φ_i is the substance concentration. Other symbols are similarly introduced. Using the splitting method one can always construct the scheme so that the presence of many components is essential only when the transformation operator is described. But if the transformation of pollutant is neglected, it will suffice to present the case of a single scalar function. The vector of model parameters is defined as follows:

$$\mathbf{Y} = (\varphi^0, \mathbf{u}, \mu, \nu, \beta, k_1, k_2, \boldsymbol{\varphi}, g_S, f_S),$$

where φ^0 is the pollutant concentration distribution at time $t = 0$.

In the numerical model the notion of parameters will be defined so that it is possible to specify not only the numerical values of the corresponding quantities but also algorithms or a collection of algorithms for them to be calculated. For instance, when the processes of pollutant transport are simulated, it is convenient to include among the model parameters algorithms for implementing the pollutant transformation operator and algorithms for calculating the elements of hydrometeorological conditions, the coefficients of turbulence exchange, pollutant sources, the coefficients of the interaction between the pollutants and the underlying surface, etc.

In order to construct discrete approximations we use a generalized formulation of the model. To this end problem (7.1) to (7.4) is assigned an integral identity. Performing a scalar multiplication of (7.1) by an arbitrary sufficiently smooth function φ^* and integrating by parts using conditions (7.2) to (7.4) we obtain an identity of the following form [7.8]:

$$\begin{aligned} I(\varphi, \mathbf{Y}, \varphi^*) = \int_{D_t} & [(\Lambda\varphi, \varphi^*) + A\varphi\cdot\varphi^* + \nu\,(\partial\varphi/\partial z)\cdot(\partial\varphi^*/\partial z) \\ & + (\mu\,\mathrm{grad}_S\,\varphi\cdot\mathrm{grad}_S\,\varphi^*) - f\cdot\varphi^*]\,dD\,dt \\ & + \int_{S_t} [(\beta\varphi\cdot\varphi^* + g_S\varphi^*)\,|_{z=z_S} + k_1\,(\varphi - \bar{\varphi})\cdot\varphi^*\,|_{z=z_H}]\,dS\,dt \\ & + \int_{\sigma_t} \left[k_2\,(\varphi - \bar{\varphi}) + \frac{1}{2}\,u_n\varphi\right]\varphi^*\,d\sigma\,dt \\ & + \frac{1}{2}\int_D \varphi\cdot\varphi^*\,|_0^t\,dD = 0, \quad (7.5) \end{aligned}$$

where $D_t = D \times [0, \bar{t}]$, $S_t = S \times [0, \bar{t}]$, $\sigma_t = \sigma \times [0, \bar{t}]$, S is a surface in D, $z = \text{const}$, and dS and $d\sigma$ the elementary areas on S and σ respectively. The expression $(\Lambda\varphi, \varphi^*)$ represents the antisymmetric form corresponding to the differential expression of the transport operator $\partial\varphi/\partial t + \mathbf{u}\ \text{grad}\ \varphi$. The basic property of this form is that for any real function defined in D_t there is an equation $(\Lambda\varphi, \varphi) = 0$. The form of this expression depends on the structure of the continuity equation. For example, for $\text{div}\ \mathbf{u} = 0$ we have

$$(\Lambda\varphi, \varphi^*) = (1/2)\,[(\partial\varphi/\partial t)\,\varphi^* - (\partial\varphi^*/\partial t)\,\varphi + \varphi^*\mathbf{u}\ \text{grad}\ \varphi - \varphi\mathbf{u}\ \text{grad}\ \varphi^*] \qquad (7.6)$$

or

$$(\Lambda\varphi, \varphi^*) = (1/2)\,[(\partial\varphi/\partial t)\,\varphi^* - (\partial\varphi^*/\partial t)\,\varphi + \varphi^*\ \text{div}\ \varphi\mathbf{u} - \varphi\ \text{div}\ \varphi^*\mathbf{u}]. \qquad (7.7)$$

It is obvious that in this case both (7.6) and (7.7) satisfy the condition

$$(\Lambda\varphi, \varphi) = 0. \qquad (7.8)$$

We take identity (7.5) as a basic for constructing discrete approximations of the model. To this end we use the finite-difference method of discretization and introduce in D_t a grid D_t^h. We approximate (7.5) on D_t^h by replacing the integrals with cubature formulas and the integrands with finite-difference relations so that in discrete representation too the symmetry and antisymmetry properties of the corresponding terms are preserved [7.8]. A discrete analogue of identity (7.5) results which is written as follows:

$$I^h(\varphi, \mathbf{Y}, \varphi^*) = 0. \qquad (7.9)$$

Here the superscript h marks discrete quantities. For notational brevity we adopt a convention that discretization operations in the functional extend to its functional arguments as well, i.e. all the functions in expression (7.9) are defined on D_t^h. The discrete analogues of problem (7.1) to (7.4) are obtained directly from discrete functional $I^h(\varphi, \mathbf{Y}, \varphi^*)$ by writing the conditions of stationarity of this functional given arbitrary and independent variations of the function φ^* at the nodes of D_t^h. Constructively these conditions are implemented using the operation of differentiation

$$\partial I^h(\varphi, \mathbf{Y}, \varphi^*)/\partial\varphi^{*h} = 0. \qquad (7.10)$$

The system of adjoint equations is similarly defined

$$\partial I^h(\varphi, \mathbf{Y}, \varphi^*)/\partial\varphi^h + \eta(\mathbf{x}, t) = 0, \qquad (7.11)$$

where $\eta(\mathbf{x}, t)$ is a function whose form is specified given particular applications of the conjugate problem. The boundary conditions in

(7.10) and (7.11) are taken into account automatically via the corresponding coefficients of discrete equations.

The introduction of conjugate problems into numerical models significantly extends the class of problems being solved and the possibilities of numerical simulation. The procedures of constructing systems of primal and adjoint problems are similar, as can be seen from (7.10) and (7.11). But these systems are functionally different. For example, the result of solving the basic system of equations is a prediction of the development in space and time of the effects of the disturbances of some factor while using the solution of the adjoint system one can calculate the functions of model sensitivity given a definite choice of functionals of quality and estimate from where disturbances will come to a given area, identify the areas of increased sensitivity of the model to the variations of parameters, design an observational experiment, estimate the model parameters, etc. [7.1], [7.3]-[7.5], [7.8].

In order to approximate identity (7.5) with respect to time discrete steps are introduced so that, given a special representation of the integrands in terms of the discrete steps, (7.10) leads to several splitting schemes.

In this case splitting "by physical processes" enables dynamic processes of transport and processes of pollutant transformation to be considered independently and the corresponding parts of the model to be implemented at different stages, joining them at the level of information exchange. Although pollutants are assumed to be multicomponent, at the transport stage each substance is expressed independently in the entire region of integration. At the transformation stage all the substances are considered simultaneously but in a localized manner at each grid point. Structurally the implementation units of the transport stage have a universal character, while those of the transformation stages are special cases for each cycle of gas phase reactions and collection of reactants. It is seen from (7.5) that the structure of the discrete identity is based on the approximations of first derivatives. Along with the conventional finite-difference schemes, central-difference ones, for example, we use implicit schemes of increased order of accuracy of rational type [7.8]. An example of this type of approximation with a three-point pattern is the scheme of a fourth order of accuracy [7.8], [7.10]:

$$\varphi'_{i+1} + 4\varphi'_i + \varphi'_{i-1} = (3/\Delta x)\,(\varphi_{i+1} - \varphi_{i-1}) + O\,(\Delta x^4), \qquad (7.12)$$

where $\varphi'_i = (\partial\varphi/\partial x)\,|_{x=x_i}$ and Δx is the grid spacing in the variable x. With a slight increase in the number of arithmetical operations these approximations have a high order of accuracy with comparatively narrow grid patterns. This is particularly important where, due to technical limitations of computers, the resolution of a discrete model

cannot be enhanced by a simple decrease of the grid spacing. Essential singularities in the use of numerical schemes for solving problems of the hydrothermodynamics of the atmosphere and those of pollutant transport arise because of the presence of regions where solutions have heavy space gradients, in the presence of localized heat, moisture, and pollutant sources, for example. The application of nonmonotonic schemes leads in such regions to undesirable computational effects (negative values of functions which by their physical meaning must be nonnegative) and to the appearance of new fictitious maxima having no physical meaning. In this connection the need arises to construct schemes having monotonicity properties and satisfying the basic conservation laws.

At present several schemes ([7.1], [7.5], [7.11], [7.12], etc.) are used to solve transport equations. The most familiar and frequently used in practice is the directed difference scheme of the first order of accuracy, but its field of application is limited by a high "computational viscosity", particularly on crude grids. Considering the existing constraints on the possibilities of enhancing the resolution of difference schemes it is natural to attempt monotonic schemes constructed on the basis of approximations of increased order of accuracy.

Quite satisfactory results are obtained when using implicit approximations of the type (7.12) with a local monotonic operator of divergence form which is introduced into the difference scheme in the range of the remainder term. For details see [7.13].

The questions concerning implementation of splitting schemes are at present extensively studied [7.1], [7.3]-[7.8]. A new element as compared to the models of the hydrothermodynamics of the atmosphere and ocean is the transformation of pollutants. At this stage within the framework of the general splitting scheme, at each point of a grid region one generally obtains systems of nonlinear algebraic equations that approximate the equations for gas phase reactions which are solved using special iterative processes taking into account the differences in the characteristic "life times" of different substances. Computing algorithms for primal and adjoint problems are similar. Slight modifications have been made for solving conjugate problems, which take into account the transposition of matrices, reverse sense of integration, reverse order of the splitting stages, and calculation of source functions via the gradients of the functionals to be evaluated.

To evaluate the quality of the models and the generalized characteristic of the processes under study, we introduce a collection of functionals defined on the set of solutions of problem (7.1) to (7.4). We shall consider functionals of two types.

(a) Functionals of the quality of a model

$$\Phi(\varphi) = \| [\varphi]_m - \varphi_m \|_{D_l^m}, \tag{7.13}$$

where φ_m and φ are the observed and calculated functions of state, respectively, and D_t^m is the set of points of D_t at which observations are made. The symbol $[\ ]_m$ denotes the operation of assigning the calculated values of φ to the set of points of D_t^m and the norm on D_t^m is given by one of the methods with a weight matrix taking into account the accuracy, reliability, and cost of observations.

(b) Functionals of a generalized estimate of the state of atmospheric pollution

$$\Phi_k(\varphi) = \int_{D_t} \eta_k(\mathbf{x}, t) \cdot Q(\varphi(\mathbf{x}, t))\, dD\, dt, \quad k = \overline{1, K}, \tag{7.14}$$

where $Q(\varphi)$ is a function of $\varphi(\mathbf{x}, t)$ and $\eta_k(\mathbf{x}, t)$ are nonnegative weight functions satisfying the conditions of normalization

$$\int_{D_t} \eta_k(\mathbf{x}, t)\, dD\, dt = 1.$$

In particular, the functions $\eta_k(\mathbf{x}, t)$ may have finite supports in D_t. Functionals of the form (7.13) are used when the quality of models is evaluated or when parameters of models are identified or when models are used to implement the space-time structure of the fields from experimental data. Functionals of the form (7.14) provide an estimate of total pollution with the weights η_k of the regions in D_t defined by the supports of weight functions.

Now consider some applications of the methods of optimization and experimental design on the basis of the given model to environmental protection problems.

7.2. EVALUATION OF MODEL SENSITIVITY TO THE VARIATIONS OF INPUT PARAMETERS

The study of the sensitivity of a model to the variations of input parameters and external effects is a necessary stage in solving problems of numerical simulation and in the study by means of numerical models of the behaviour of real physical systems [7.8], [7.14]. By definition sensitivity functions are derivatives of the characteristics under study with respect to the parameters of models. It is convenient to take as a measure of sensitivity one of the functionals of the type (7.13) or (7.14) given on the set of state functions. In this case, using formally the methods of variational calculus to calculate the variations of the functional $\delta\Phi_k(\varphi)$ and the sensitivity functions $\operatorname{grad}_{\mathbf{Y}} \Phi_k(\varphi)$ yields the constructive formula [7.8]

$$\delta\Phi_k(\varphi) \equiv (\operatorname{grad}_{\mathbf{Y}} \Phi_k, \delta\mathbf{Y}) = \frac{\partial}{\partial\xi} [I^h(\varphi, \mathbf{Y} + \xi\delta\mathbf{Y}, \varphi^*)]\,|_{\xi=0}, \tag{7.15}$$

where ξ is a real parameter, $\delta\mathbf{Y}$ is the vector made up of the variations of parameters, φ the solution of the fundamental system of equations (7.10) for a given vector of parameters $\mathbf{Y}$, and φ^* the solution of the conjugate problem (7.11) with $\varphi^* \mid_{t=\bar{t}} = 0$ and the source function

$$\eta(\mathbf{x}, t) \equiv \partial\Phi_k^h(\varphi)/\partial\varphi^h. \tag{7.16}$$

The algorithm for constructing the adjoint system of equations (7.11) reduces to performing the following operations:

$$\frac{\partial}{\partial\varphi'^h}\left\{\frac{\partial}{\partial\xi}\left[I^h(\varphi+\xi\varphi', \mathbf{Y}, \varphi^*)+\Phi_k(\varphi+\xi\varphi')\right]\mid_{\xi=0}\right\}=0. \tag{7.17}$$

For model (7.1) to (7.4) the basic relation of sensitivity (7.15) is a discrete analogue of the expression

$$\begin{aligned}\delta\Phi_k(\varphi, \mathbf{Y}) = &\int_{D_t}\left[(\delta A\varphi\cdot\varphi_k^*)+\delta\mathbf{u}(\varphi_k^*\operatorname{grad}\varphi-\varphi\operatorname{grad}\varphi_k^*)\right.\\ &\left.+\delta\nu(\partial\varphi/\partial z)(\partial\varphi_k^*/\partial z)+\delta\mu\operatorname{grad}_S\varphi\cdot\operatorname{grad}_S\varphi_k^*-\delta f\varphi_k^*\right]dD\,dt\\ &+\int_{S_t}\left[(\delta\beta\varphi\cdot\varphi_k^*+\delta g_S\varphi_k^*)\mid_{z=z_S}+(\delta k_1(\varphi-\bar{\varphi})\cdot\varphi_k^*-k_1\delta\bar{\varphi}\cdot\varphi_k^*)\mid_{z=z_H}\right]dS\,dt\\ &+\int_{\sigma_t}\left\{\left[\delta k_2(\varphi-\bar{\varphi})-k_2\delta\bar{\varphi}\right]\varphi_k^*+\frac{1}{2}\delta u_n\varphi\cdot\varphi_k^*\right\}d\sigma\,dt\\ &+\frac{1}{2}\int_D\delta\varphi\cdot\varphi_k^*\mid_{t=0}dD.\end{aligned} \tag{7.18}$$

Formulas (7.15) and (7.18) show clearly the physical meaning of the functions of model sensitivity. They are equal to the coefficients of the corresponding variations of parameters and for them to be calculated requires solving simultaneously the primal and adjoint problems. It should be noted that the primal problem is solved only once whereas adjoint problems are solved as many times as there are functionals for the study of model behaviour and that all these problems are solved for unperturbed values of parameters. The sensitivity functions are obtained from the expression for the discrete functional by differentiation with respect to the parameter variations, i.e.

$$\operatorname{grad}_{\mathbf{Y}}\Phi_k(\varphi)\equiv\frac{\partial}{\partial\delta\mathbf{Y}}\left\{\frac{\partial}{\partial\xi}\left[I^h(\varphi, \mathbf{Y}+\xi\delta\mathbf{Y}, \varphi^*)\right]\mid_{\xi=0}\right\}. \tag{7.19}$$

For examples of calculating sensitivity functions for a model of the type (7.1) to (7.4) on a sphere by the above algorithm see [7.15]. When the sensitivity of a model is investigated the conjugate function plays the part of the generalized Lagrangian multiplier to take

into account constraints in the form of the equations of model (7.1) to (7.4). In [7.1], [7.4], [7.5] the solution of the adjoint problem is interpreted as the function of the influence of the parameters of the model for linear functionals of the form (7.14). To show how the conjugate function plays this part we write the formula for evaluating functionals (7.14) with $Q(\varphi) = \varphi(\mathbf{x}, t)$ and with a transformation linear operator directly via the input parameters of the model

$$\Phi_k(\varphi) = \int_{D_t} f(\mathbf{x}, t) \cdot \varphi_k^*(\mathbf{x}, t)\, dD\, dt + \int_{S_t} (g_S \varphi_k^* |_{z=z_S} + k_1 \bar{\varphi} \varphi_k^* |_{z=z_H})\, dS\, dt + \int_{\sigma_t} k_2 \bar{\varphi} \varphi_k^*\, d\sigma\, dt + \int_D \varphi_k^* \cdot \varphi^0 |_{t=0}\, dD, \quad (7.20)$$

where φ_k^* is obtained from the solution of the adjoint problem (7.11) with a source term $\eta(\mathbf{x}, t) \equiv \operatorname{grad}_\varphi \Phi_k(\varphi) = \eta_k(\mathbf{x}, t)$ and initial conditions $\varphi_k^* |_{t=\bar{t}} = 0$.

Thus using sensitivity functions implements the relation of the variations of parameters and estimated functionals and relations of the type (7.15), (7.18), and (7.20) provide working formulas for a qualitative analysis of numerical models. Physically these formulas contain information about the relative contribution of the variations of different parameters to the variations of functionals and computationally formulas (7.15) and (7.18) provide an error estimate for the desired characteristic if the input data contain errors $\delta \mathbf{Y}$. These may be determinate or random. In solving experimental design problems the function φ^* determines by (7.15) and (7.18) the location of measuring devices which is optimal in that the error $\delta\Phi_k(\varphi)$ of the measured functional $\Phi_k(\varphi)$ is independent of the errors of the state function.

7.3. IMPLEMENTATION OF THE SPACE-TIME STRUCTURE OF POLLUTANT CONCENTRATION FIELDS FROM OBSERVED DATA

The essence of the problem is to estimate the current state of the pollution of the atmosphere or water objects by using simultaneously the observed data and the results of mathematical modelling [7.8]. Particularly helpful are methods of estimating the current state of the object under investigation using models in the regime of utilizing actual data where there is a lack of measuring devices or the cost of individual observations is too great and the

observations themselves are heterogeneous in character, asynchronous in time, and irregularly located in space. Let Ψ_{obs} be a set of observed values and $\Psi(\varphi)$ be a collection of operators which describes implementation of observations on a set $D_t^m \in D_t^h$.

Then the quality functional may be introduced as follows:

$$J_0(\varphi) = \| \Psi_{obs} - \Psi(\varphi) \|_{D_t^m}. \tag{7.21}$$

The form of the operators $\Psi(\varphi)$ is determined by the functional content of observations complexes, these may be direct and indirect observations, and the norm in (7.21) is introduced so that it is possible to consider prior information about the accuracy, value, and cost of the observations by using the corresponding weights. Extra terms containing prior information about the input parameters of the model are introduced with some weights into the quality functional. These terms are used to regularize the process of finding an optimum solution.

The quality functional is chosen so that it is continuous and differentiable with respect to the components of the state function, i.e. the function

$$\mathrm{grad}_\varphi J_0(\varphi) \equiv \partial J_0(\varphi)/\partial\varphi \tag{7.22}$$

should exist and be defined on the range of the state function at the nodes of the grid region D_t^h. For example, if the model is discretized on a regular grid $D_t^h \in D_t$ and only direct observations of the values of functions φ on an irregular collection of points of D_t^m are considered. then $\Psi_{obs} = \varphi_{obs}$, where φ_{obs} are observed values of the state function, and one may take as $\Psi(\varphi)$ an operator of the interpolation of simulated functions of state from a regular grid D_t^h to an irregular grid D_t^m in which φ_{obs} are given.

It is thus required to solve the problem of minimizing the quality functional (7.21) on the set of allowed values of parameters $\mathbf{Y}$ provided the parameters and the state function are related by the discrete analogue (7.10) of model (7.1) to (7.4).

To solve this problem we use iteration methods of gradient type. The application of the splitting method to the solution of systems of basic and adjoint equations allows a simple and efficient scheme of implementing the problem in spite of the large size of the problem. Adaptation of data will be effected in the time interval $[0, \bar{t}]$ drifting in some reference frame as information comes in. The length of the interval is a parameter of the model. Among the desired parameters is the value of the state function at time $t = 0$.

The scheme of the computing algorithm is as follows:

1. Specify the value of the vector of input parameters $\mathbf{Y}^0$.
2. Solve the system of basic equations (7.10) in the interval

$[0, \bar{t}]$ for the given value of parameters $\mathbf{Y}^{(n)}$, $n = 0, 1, \ldots$. A solution $\varphi^{(n)} = \varphi(\mathbf{x}, t, \mathbf{Y}^{(n)})$ results.

3. Solve the adjoint problems (7.11) in the interval $[\bar{t}, 0]$ subject to the conditions

$$\varphi^* |_{t=\bar{t}} = 0, \quad \eta(\mathbf{x}, t) \equiv \partial J_0^h(\varphi)/\partial \varphi^h. \tag{7.23}$$

4. Calculate the sensitivity functions of the functional (7.21) via the solutions of the basic and adjoint equations from the formula

$$\operatorname{grad}_{\mathbf{Y}} J_0(\varphi) = \frac{\partial}{\partial \delta \mathbf{Y}} \left[\frac{\partial}{\partial \xi} I^h(\varphi^{(n)}, \mathbf{Y}^{(n)} + \xi \delta \mathbf{Y}, \varphi^{*(n)}) \right] \Big|_{\xi=0}, \tag{7.24}$$

where $\varphi^{(n)}$ and $\varphi^{*(n)}$ are solutions of problems (7.10) and (7.11), (7.23) for $\mathbf{Y} = \mathbf{Y}^{(n)}$. In particular, the functions of the sensitivity of the functional $J_0(\varphi)$ to the variations of the initial value of the state function φ^0 are given as follows:

$$\operatorname{grad}_{\varphi^0} J_0(\varphi) = \varphi^* |_{t=0}. \tag{7.25}$$

5. Find a new value of the vector of parameters from the formula

$$\mathbf{Y}^{(n)} = \mathbf{Y}^{(n-1)} - \varkappa_n \Gamma \operatorname{grad}_{\mathbf{Y}} J_0(\varphi) |_{\mathbf{Y}=\mathbf{Y}^{(n-1)}}, \tag{7.26}$$

where $\varkappa_n$ is the iteration parameter which is determined from the condition of the minimum of the functional $J_0^h(\varphi)$ in the direction of the vector of parameters $\mathbf{Y}^{(n)}$ and Γ is a nonnegative matrix (possibly a diagonal one) giving the number of varied parameters and the structure of the new approximation of $\mathbf{Y}^{(n)}$. For instance, in the problem of reconstructing the fields the only parameter varied is φ^0. The choice of Γ is determined by the investigator when solving a particular problem. The algorithm for an approximate estimation of the iteration parameter $\varkappa_n$ is described in [7.8]. With the new value of $\mathbf{Y}^{(n)}$ the computational cycle is repeated until a given criterion of convergence, which considers the magnitude of changes in the parameter vector and the quality functional [(7.10)], is satisfied. It is obvious that it is sufficient to attain the convergence of iterations only within the accuracy stipulated by the accuracy of observations and errors of the model itself.

The iterative process generally converges to some local minimum in the neighbourhood of a given initial value of the vector $\mathbf{Y}^0$ and an a priori estimate of the desired characteristics considered in the functional-regulator, so it is desirable to use as much prior information as possible to facilitate the search process.

The implementation of the above procedure comprises two stages, prediction and adaptation. This makes it very convenient for practical application in a system of monitoring the quality of the atmosphere, particularly where there are not enough observational stations to cover an entire territory but there are no hard constraints

on the number of observations in time. For particular examples of applying optimization methods to reconstruct fields see [7.8], [7.16], [7.17].

7.4. EXPERIMENTAL DESIGN IN ENVIRONMENTAL PROTECTION PROBLEMS

Methods of the theory of experimental design are widely used in practical problems associated with the optimization of conditions and costs of making experiments with a view to evaluating the parameters of mathematical models and the real physical objects described by these models with a maximum attainable accuracy. General principles of experimental design theory have been developed by now, but their implementation and the meaning of the notion of "optimum experiment", as well as the optimality criteria, depend on a particular statement of a problem [7.5], [7.18], [7.19]. We now consider a method of solving inverse problems in the estimation of parameters of the pollutant source distribution in the atmosphere on the basis of the pollutant transport model (7.1) to (7.4) [7.20]. Suppose the pollutant transformation operator is linear in concentrations and the source functions $f(\mathbf{x}, t)$ and $g_S(\mathbf{x}, t)$ in (7.1) and (7.2) contain the desired parameters $\boldsymbol{\lambda}$ and $\mathbf{q}$, i.e.

$$f \equiv f(\mathbf{x}, t, \boldsymbol{\lambda}), \quad g_S \equiv g_S(\mathbf{x}, t, \mathbf{q}). \tag{7.27}$$

To find the function $\varphi(\mathbf{x}, t, \boldsymbol{\lambda}, \mathbf{q})$ and vectors $\boldsymbol{\lambda}$ and $\mathbf{q}$ we give the following information about the actual pollutant distribution:

$$\zeta_i(t) = \varphi(\mathbf{x}_i, t, \boldsymbol{\lambda}, \mathbf{q}) + \gamma_i \eta(t), \quad 0 \leqslant t \leqslant \bar{t}, \quad \mathbf{x}_i \in D, \tag{7.28}$$

$$\mathbf{E}[\gamma_i] = 0, \quad \mathbf{E}[\gamma_i \gamma_j] = \delta_{ij} R^{-1}(\mathbf{x}_i), \quad i, j = \overline{1, n}.$$

Here n is the number of points $\mathbf{x}_i$ at which observations are made, $\mathbf{E}$ the expectation operator, and δ_{ij} the Kronecker delta.

The functions $\zeta_i(t)$, $\eta(t)$, and $R(\mathbf{x})$ are given, the value of $\varphi_i(t)$ is the result of the observation of the pollutant concentration at a point $\mathbf{x}_i$ at time t, the function $R(\mathbf{x})$ characterizes the effectiveness of observations at the points of a region Ω, and $\eta(t)$ allows for the specific character of the observations in time.

We denote by N the total number of observations and by r_i the number of observations at a point $\mathbf{x}_i$ ($\sum_{i=1}^{n} r_i = N$) and introduce the following definition.

Definition [7.21]. *The collection of quantities*

$$\varepsilon_N = \{\mathbf{x}_i,\ p_i \mid p_i = r_i N^{-1},\quad i = \overline{1,\ n}\} \tag{7.29}$$

is called a design of an experiment.

By a solution of the inverse problem we shall mean the estimates $\hat{\boldsymbol{\lambda}}$ and $\hat{\mathbf{q}}$ of the least squares method (LSM estimates) for unknown parameters $\boldsymbol{\lambda}$ and $\mathbf{q}$ and the function $\hat{\varphi}_N = \varphi(\mathbf{x},\ t,\ \hat{\boldsymbol{\lambda}},\ \hat{\mathbf{q}})$. Solution in this sense is introduced in [7.22]. LSM estimates minimize the functional

$$J_N(\boldsymbol{\lambda},\ \mathbf{q}) = \sum_{i=1}^{n} p_i R(\mathbf{x}_i) \int_0^{\overline{t}} [\zeta_i(\tau) - \varphi(\mathbf{x}_i,\ \tau,\ \boldsymbol{\lambda},\ \mathbf{q})]^2 d\tau \tag{7.30}$$

subject to the condition that the function $\varphi(\mathbf{x},\ \tau,\ \boldsymbol{\lambda},\ \mathbf{q})$ is in the set of solutions of the discrete analogue (7.10) of problem (7.1) to (7.4). The subscript N indicates that the estimates correspond to observations (7.29).

Let the functions $f(\mathbf{x},\ t,\ \boldsymbol{\lambda})$ and $g_S(\mathbf{x},\ t,\ \mathbf{q})$ be represented as

$$f(\mathbf{x},\ t,\ \boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k f_k(\mathbf{x},\ t),\quad g_S(\mathbf{x},\ t,\ \mathbf{q}) = \sum_{m=1}^{P} q_m g_m(\mathbf{x},\ t), \tag{7.31}$$

where $\{f_k\}_{k=1}^{K}$ and $\{g_m\}_{m=1}^{P}$ are linearly independent systems of functions. Then the solution of the primal problem can be written as

$$\varphi(\mathbf{x},\ t,\ \theta) = \Phi(\mathbf{x},\ t) + \sum_{m=1}^{M} \theta_m U_m(\mathbf{x},\ t) \tag{7.32}$$

where $M = K + P$,

$$\theta_m = \begin{cases} \lambda_m, & m = \overline{1,\ K}, \\ q_{m-K}, & m = \overline{K+1,\ M}, \end{cases}$$

and the functions $U_m(\mathbf{x},\ t)$ and $\Phi(\mathbf{x},\ t)$ are solutions of a set of problems:

$$\begin{gathered} lU_k = f_k(\mathbf{x},\ t),\quad LU_m = 0,\quad L\Phi = 0,\quad lU_k = 0, \\ lU_m = g_{m-K}(\mathbf{x},\ t),\quad l\Phi = 0 \text{ for } z = z_S, \\ \nu \frac{\partial U_k}{\partial z} + k_1 U_k = 0,\quad \nu \frac{\partial U_m}{\partial z} + k_1 U_m = 0, \\ \nu \frac{\partial \Phi}{\partial z} = -k_1(\Phi - \bar{\varphi}) \quad \text{for } z = z_H, \end{gathered} \tag{7.33}$$

$$\mu \frac{\partial U_k}{\partial n} + k_2 U_k = 0, \quad \mu \frac{\partial U_m}{\partial n} + k_2 U_m = 0,$$

$$\mu \frac{\partial \Phi}{\partial n} = -k_2 (\Phi - \bar{\varphi}) \text{ on } \sigma, \quad (7.33)$$

$$U_k = 0, \quad U_m = 0, \quad \Phi = \varphi^0(x) \text{ for } t = 0, \quad k = \overline{1, K},$$

$$m = \overline{K+1, M}.$$

The discrete approximations and the algorithms for solving these problems are the same as those for the primal problem (7.1) to (7.4) and are defined by scheme (7.10).

It follows from (7.30), (7.32) that LSM estimates θ are found from the formula [7.21]

$$\hat{\theta} = M_*^{-1}(\varepsilon)\, \mathbf{V}, \quad (7.34)$$

where $M_*(\varepsilon)$ is a square $M \times M$ matrix with components

$$M_{ks}(\varepsilon) = \sum_{i=1}^{n} p_i R(\mathbf{x}_i) \int_0^{\bar{t}} U_k(\mathbf{x}_i, \tau) \cdot U_s(\mathbf{x}_i, \tau)\, d\tau, \quad (7.35)$$

$\mathbf{V}$ is a vector of dimension M,

$$V_k = \sum_{i=1}^{n} p_i R(\mathbf{x}_i) \int_0^{\bar{t}} U_k(\mathbf{x}_i, \tau) [\zeta_i(\tau) - \Phi(\mathbf{x}_i, \tau)]\, d\tau, \quad k, s = \overline{1, M}, \quad (7.36)$$

and the normalized variance matrix $D(\varepsilon)$ of estimates $\hat{\theta}$ is equal to $[M_*(\varepsilon)]^{-1}$.

It is common practice to use convex functionals F dependent on $D(\varepsilon)$ as an accuracy measure of a solution of inverse problems. The experimental design problem consists in seeking a design ε^* for which

$$F[D(\varepsilon^*)] = \inf_{\varepsilon_N} F[D(\varepsilon_N)]. \quad (7.37)$$

A direct solution of problem (7.37) is difficult for large M, so one should apply special methods for seeking optimum designs. For definiteness we discuss ways of constructing D-optimum designs, i.e. designs minimizing the value of the determinant of the variance matrix $D(\varepsilon)$.

Consider the following iteration procedure [7.23].

(1) Specify a nonsingular design ε_0 (the initial estimate)

$$\varepsilon_0 = \{\mathbf{x}_i, p_i, i = \overline{1, n}\}, \quad n \geqslant M. \quad (7.38)$$

(2) Calculate the elements of the information matrix $M_*(\varepsilon)$, from formula (7.35), and its inverse, the variance matrix $D(\varepsilon_0)$.

(3) Find the point $\mathbf{x}_0 \in \Omega$, at which

$$\max_{\mathbf{x}\in\Omega} R(\mathbf{x}) \cdot d(\mathbf{x}, \varepsilon_0),$$

and the corresponding value of $R(\mathbf{x}_0)\, d(\mathbf{x}_0, \varepsilon_0)$ where

$$d(\mathbf{x}, \varepsilon) = \int_0^{\bar{t}} \mathbf{U}'(\mathbf{x}, t)\, D(\varepsilon)\, \mathbf{U}(\mathbf{x}, t)\, dt.$$

(4) Compile a design

$$\varepsilon_1 = (1 - \alpha_0)\, \varepsilon_0 + \alpha_0 \varepsilon(\mathbf{x}_0), \tag{7.39}$$

where $\varepsilon(\mathbf{x}_0)$ is a design consisting of a single point $\mathbf{x}_0$.

The value of α_0 is chosen in such a way that the increase in the information matrix should be maximal, i.e.

$$\alpha_0 = \delta_0/[M(\delta_0 + M - 1)], \tag{7.40}$$

where $\delta_0 = R(\mathbf{x}_0)\, d(\mathbf{x}_0, \varepsilon_0) - M$.

Operations (7.2) to (7.4) are then repeated, with ε_0 replaced by ε_1, etc. The iterative process (7.1) to (7.4) converges ([7.21], [7.23]), with

$$\lim_{s\to\infty} |M_*(\varepsilon_S)| = |M_*(\varepsilon^*)|,$$

where $M_*(\varepsilon^*)$ is an information matrix corresponding to a D-optimum design and $M_*(\varepsilon_S)$ is a matrix corresponding to a design ε_S.

For examples of solving design problems see [7.20]. For nonlinear models an iteration scheme of implementing the experimental design procedure with nonlinearity iterations is constructed.

In practice there is usually prior information about the capacities of pollutant sources and the hydrometeorological conditions in the areas involved. Therefore, when solving experimental design problems it is necessary to use this information to a maximum extent, i.e. a fixed network of observational stations should be constructed for the most general meteorological situation and a running system set up according to the specific conditions by using the technique of sequential planning together with the results of numerical simulation of pollutant concentrations in the atmosphere.

Thus, methods of experimental design make it possible to estimate the capacities of the sources polluting the atmosphere. From the point of view of environmental protection this means that the use of mathematical models in the system of monitoring the quality of the atmosphere allows industrial installations breaking the permissible standards of pollution to be detected to a considerable degree of reliability from observed data on pollutant concentrations and the results of simulation. But from the practical point of view it is better to observe directly the intensities of pollutant ejections

and to enter this information into the models to be used for the analysis of the processes of transport and transformation of pollutants under varied hydrometeorological conditions. By common sense and economically such a strategy is more advantageous. Knowing the distribution and capacities of the sources it is possible to formulate from the results of predicting the quality of the atmosphere a program of optimal control of the intensity of ejections proceeding from the standards of permissible loads on the environment and the cost of the damage caused by pollutions to the natural environment together with the expenses of implementing measures for the reduction of ejections. Problems of this type are considered in the next section.

7.5. OPTIMIZATION MODELS OF THE OPERATING CONDITIONS OF SOURCES POLLUTING THE ATMOSPHERE

Models for the control of the quality of the natural environment are a necessary working tool in handling questions of hydrometeorological and ecological justification of economic projects and in planning nature conservation measures that require the control over the ejections of operating industrial installations with allowance for the special features of hydrometeorological conditions and constraints of sanitary and socioeconomic character. A number of mathematical models for solving this kind of problems are formulated in [7.1], [7.24]-[7.28]. It is obvious that all constraints must be considered taken together since along with actual expenses of implementing the projects an optimum solution must consider the price of possible consequences of the implementation to the environment.

Some optimization problems of pollutant ejections from a collection of point, linear, plane, or volume sources [7.29] are considered below. The requirement that the cost of reducing the ejections to a permissible level should be minimal is taken as a criterion. Formulation of cost functions takes into account the damage caused by pollution and the cost of measures of technological character for reducing the ejection. Also considered is the case where the available economic resources are insufficient for a permissible level to be attained and the condition that the pollution of the subregions involved should be minimal is taken as a criterion.

In discussing fundamental questions in the statements of the problems and approaches to their solution we restrict our consideration to a linear version of the model for a total pollutant concentration. For convenience of further presentation we specify the meaning of the parameters λ and $\mathbf{q}$ in (7.27). Suppose that in a region D there are K pollutant point sources located at points $\mathbf{x}_i$ $(i = \overline{1, K})$ with the intensities of ejections λ_i $(i = \overline{1, K})$, respectively, and that

on a surface S, with $z = z_S$, there are P sources a_i $(i = \overline{1, P})$ that form a collection of point, linear, and plane sources with intensities q_i $(i = \overline{1, P})$. Then the right-hand sides of (7.1), (7.2) can be represented as

$$f(\mathbf{x}, t) = \sum_{i=1}^{K} \lambda_i \delta(\mathbf{x} - \mathbf{x}_i), \quad g_S(\mathbf{x}, t) = \sum_{i=1}^{P} q_i g_i(\mathbf{x}, t), \quad (7.41)$$

where $\delta(\mathbf{x} - \mathbf{x}_i)$, $i = \overline{1, K}$, is the delta function concentrated at a point $\mathbf{x}_i$ and $g_i(\mathbf{x}, t)$ is the distribution function of sources on the Earth's surface. By the assumptions made the pollutant concentration $\varphi(\mathbf{x}, t)$ may be determined from (7.32). The meaning of the solution becomes clear if one compares expressions (7.31) and (7.41). This formula remains valid in the general case when the source functions f and g_S are represented as linear combinations of arbitrary systems of functions dependent on space coordinates and time, with only the form of the right-hand sides in the problems of finding the fundamental system of solutions changed, the algorithm remaining invariable.

We introduce the following notation: ω_m is the reduction in the intensity of the pollutant ejection of the source located at a point $\mathbf{x}_m$, $m = \overline{1, K}$, and of the source a_{m-K} with $m = \overline{K+1, M}$; θ_m (ω_m/θ_m) the function characterizing the cost of the reduction in the ejection intensity by ω_m of the corresponding source, and Ω_m the maximum reduction in the ejection intensity of the mth source, $0 \leqslant \Omega_m \leqslant \theta_m$.

The cost functions are chosen so that it is possible to consider simultaneously the cost of the damage caused by the environmental pollution and the expenses of reducing the volume of production or improving the technology and sewage purification works required for the ejection intensity to be reduced to a certain level [7.1], [7.30]. The total function of the cost of impurity ejection control on an industrial area scale is defined as the sum of costs of the regulation of individual sources.

If the intensity of the impurity ejection from each source is reduced by ω_m, $m = \overline{1, M}$, respectively, the pollutant concentration, by the linearity of the problem, is defined as

$$q(\mathbf{x}, t, \mathbf{l}) = \Phi(\mathbf{x}, t) + \sum_{m=1}^{M} \theta_m (1 - l_m) U_m(\mathbf{x}, t), \quad (7.42)$$

where $\mathbf{l} = (l_1, \ldots, l_M)$, $l_m = \omega_m/\theta_m$.

Suppose that in D there are several subregions σ_k, $k = \overline{1, s}$, in which the function $\varphi(\mathbf{x}, t, \mathbf{l})$ is subject to the following con-

straints [7.1], [7.24]:

$$J_k(q) = \int_0^{\bar{t}} \int_{\sigma_k} \eta_k(\mathbf{x}, t)\, \varphi(\mathbf{x}, t, \mathbf{l})\, d\mathbf{x}\, dt \leqslant F_k \quad k = \overline{1, s}, \tag{7.43}$$

where F_k are given numbers and $\eta_k(\mathbf{x}, t)$ are nonnegative weight functions satisfying the normalization conditions $\int_0^{\bar{t}} \int_D \eta_k(\mathbf{x}, t)\, d\mathbf{x}\, dt = 1$. Considering (7.42) conditions (7.43) assume the form

$$\sum_{m=1}^{M} r_{km} l_m \geqslant \Phi_k, \tag{7.44}$$

where

$$r_{km} = \theta_m \int_0^{\bar{t}} \int_{\sigma_k} \eta_k(\mathbf{x}, t)\, U_m(\mathbf{x}, t)\, d\mathbf{x}\, dt,$$

$$\Phi_k = \int_0^{\bar{t}} \int_{\sigma_k} \eta_k(\mathbf{x}, t)\, \varphi(\mathbf{x}, t, \mathbf{l})\, |_{\mathbf{l}=0}\, d\mathbf{x}\, dt - F_k,$$

$$k = \overline{1, s}; \quad m = \overline{1, M}.$$

Problem 1. Find on a set E control $\mathbf{l}_*$ minimizing the function

$$V(\mathbf{l}) = \sum_{m=1}^{M} G_m(l_m) \tag{7.45}$$

subject to

$$\sum_{m=1}^{M} r_{km} l_m \geqslant \Phi_k, \quad k = \overline{1, s}, \tag{7.46}$$

where

$E = \{\mathbf{l}: 0 \leqslant l_m \leqslant E_m \leqslant 1, \quad m = \overline{1, M}\}, \quad E_m = \Omega_m/\theta_m.$

If $G_m(l_m)$ are defined and convex in $[0, E_m]$, then any point of a local minimum of the function $V(\mathbf{l})$ is simultaneously a point of the global minimum. Suppose in addition that $G_m(l_m)$ are continuously differentiable in the corresponding intervals and the set of points of the minimum E_* is nonempty. Then for every point $\mathbf{l}_* \in E_*$ there are Lagrangian multipliers $\lambda^* = (\lambda_1^*, \ldots, \lambda_s^*)$, $\lambda_i^* \geqslant 0$, $i = \overline{1, s}$, such that the pair $(\mathbf{l}_*, \lambda^*)$ forms a saddle point of the Lagrangian [7.31].

It is convenient to represent a solution in the form (7.42) in the following cases:

(i) where it is required to control the sources so that the pollution at any point of D does not exceed a specified level;

(ii) where a general structure of source functions is given or a source distribution in space is fixed and where the source parameters to be controlled are independent of time;

(iii) where the number of desired parameters does not exceed the number of constraints, i.e. $M \leqslant s$.

Any linear functional of the solution of a problem of the form (7.1) to (7.4) can be calculated directly via the nonhomogeneity of the right-hand side and the boundary conditions using the solution of the adjoint problem. Let $w = 0$ for $z = z_s$ and $z = z_H$. Then for the functional $J_k(\varphi)$ and using the notation of (7.20) we get

$$J_k(\varphi) = \int_{D_t} \varphi_k^*(\mathbf{x}, t) \cdot \sum_{m=1}^{K} \theta_m (1 - l_m)\, \delta(\mathbf{x} - \mathbf{x}_m)\, dD\, dt$$

$$+ \int_D \varphi_k^*(\mathbf{x}, t) \cdot \varphi^0(\mathbf{x})\,|_{t=0}\, dD$$

$$+ \int_{s_t} (\varphi_k^*)\,|_{z=z_s} \sum_{m=K+1}^{M} \theta_m (1 - l_m) + g_{m-K}(\mathbf{x}))\, ds\, dt$$

$$+ \int_{s_t} k_1 \bar{\varphi} \varphi_k^*\, ds\, dt\,|_{z=z_H} + \int_{\sigma_t} k_2 \bar{\varphi} \cdot \varphi_k^*\, d\sigma\, dt, \qquad (7.47)$$

where $\varphi_k^*(\mathbf{x}, t)$, $k = \overline{1, s}$, are solutions of the adjoint problems (7.11) subject to the following conditions:

$$\varphi_k^*(\bar{t}) = 0, \quad \eta(\mathbf{x}, t) \equiv \eta_k(\mathbf{x}, t). \qquad (7.48)$$

Both formulas for the calculation of the functionals $J_k(q)$, $k = \overline{1, s}$, are equivalent with the above constraints on the form of the right-hand sides. The functions $\varphi_k^*(\mathbf{x}, t)$ play the part of Lagrangian multipliers to allow for the discrete analogues of equations (7.1) to (7.4) regarded as constraints on the set of functions $\varphi(\mathbf{x}, t)$. For representations (7.47) to be constructively implemented it is required to solve s adjoint problems (7.11), (7.48). Therefore, if the condition $s < M$ is met, then the computing algorithm for determining (7.47) is more economical than that for (7.43) allowing for (7.42).

The principal merit of formulas (7.47) is that they are true for any representation of the source functions and initial field $\varphi^0(\mathbf{x})$.

Using (7.47) the coefficients r_{km} and Φ_k in constraints (7.44),

(7.46) assume the form

$$r_{km} = \begin{cases} \theta_m \int\limits_0^{\bar{t}} \varphi_k^*(\mathbf{x}_m, t)\, dt, & m = \overline{1, K}, \\ \theta_m \int\limits_{s_t} \varphi_k^*(\mathbf{x}, t)\, g_{m-K}(\mathbf{x})\,|_{z=z_s}\, ds\, dt, & m = \overline{K+1, M}, \end{cases} \tag{7.49}$$

$$\Phi_k = \int\limits_0^{\bar{t}} \sum_{m=1}^{K} \theta_m \varphi_k^*(\mathbf{x}_m, t)\, dt + \sum_{m=K+1}^{M} \int\limits_{s_t} \theta_m q_k^*(\mathbf{x}, t)$$

$$+ g_{m-K}(\mathbf{x})\,|_{z=z_s}\, ds\, dt + \int\limits_{s_t} k_1 \bar{\varphi} \cdot \varphi_k^*\,|_{z=z_H}\, ds\, dt$$

$$+ \int\limits_{\sigma_t} k_2 \bar{\varphi} \varphi_k^*\, d\sigma\, dt + \int\limits_D \varphi_k^*(\mathbf{x}, t) \cdot \varphi^0(\mathbf{x})\,|_{t=0}\, d\mathbf{x} - F_k, \quad k = \overline{1, s}. \tag{7.50}$$

With such a formulation of constraints it is possible to minimize the cost function (7.45) not only by the choice of parameters of source control but also by a suitable location of the sources. Since no information about the distribution and structure of the sources is required for the adjoint problems (7.11), (7.48) to be solved, constraints of the form (7.44), (7.49), (7.50) make it possible to identify feasible areas for the location of sources given any set of values of control parameters. The algorithm for solving such a problem is proposed in [7.1], [7.24].

Suppose now that a resource $V(\mathbf{l})$ is bounded above by V_0 and constraint (7.46) holds for no vector $\mathbf{l}$ on the set $E_0 = \{\mathbf{l} \in E,\ V(\mathbf{l}) \leqslant V_0\}$. Then it is appropriate to consider the following problem.

Problem 2. Minimize on a set E the functional

$$I(\mathbf{l}) = \sum_{k=1}^{s} \rho_k \left[F_k - \int\limits_t^{t+\bar{t}} \int\limits_D \eta_k(\mathbf{x}, t)\, \varphi(\mathbf{x}, t, \mathbf{l})\, d\mathbf{x}\, dt \right]^2 \tag{7.51}$$

subject to

$$\sum_{m=1}^{M} G_m(l_m) \leqslant V_0, \tag{7.52}$$

where ρ_k are given nonnegative weights.

In the case $s = 1$ the problem of minimizing functional (7.51) consists in minimizing on E subject to (7.52) the functional

$$I(\mathbf{l}) = \int\limits_t^{t+\bar{t}} \int\limits_\sigma \eta(\mathbf{x}, t)\, \varphi(\mathbf{x}, t, \mathbf{l}), \quad \sigma \in D \tag{7.53}$$

Suppose the functions G_m (l_m) are convex, the set of points of the minimum of the functional I (l) is nonempty, and there is a point $\mathbf{l} \in E$ such that

$$\sum_{m=1}^{M} G_m(\hat{l}_m) < V_0. \tag{7.54}$$

Then for every point $\mathbf{l}_* \in E_*$ there is a number λ^*, $\lambda^* \geqslant 0$, such that the pair $(\mathbf{l}_*, \lambda^*)$ forms a saddle point of the Lagrangian [7.31].

Consider now the case where it is appropriate to vary the ejection intensity of given pollutant sources in time as a function of the meteorological conditions, for example. Then the coefficients q_n, $n = \overline{1, P}$, and λ_k, $k = \overline{1, K}$, appearing in equation (7.1) and the boundary condition (7.2) will depend on $t \in [0, \bar{t}]$.

Problem 3. Suppose in the region D_t the function φ (x, t) represents a solution of problem (7.1) to (7.4) for

$$f = \sum_{n=1}^{K} \lambda_n(t)\, \delta(\mathbf{x} - \mathbf{x}_n), \quad g_s = \sum_{k=1}^{P} q_k(t)\, g_k(\mathbf{x}), \tag{7.55}$$

where $\boldsymbol{\lambda}(t) = (\lambda_1(t), \ldots, \lambda_K(t))$, $\mathbf{q}(t) = (q_1(t), \ldots, q_P(t) \in L_2(0, T)$,

$$\int_0^{\bar{t}} \int_{\sigma} \varphi(\mathbf{x}, t, \boldsymbol{\lambda}, \mathbf{q})\, d\mathbf{x}\, dt \leqslant C = \text{const}, \tag{7.56}$$

$$0 \leqslant \boldsymbol{\lambda}^{(1)} \leqslant \boldsymbol{\lambda}(t) \leqslant \boldsymbol{\lambda}^{(2)}, \quad 0 \leqslant \mathbf{q}^{(1)} \leqslant \mathbf{q}(t) \leqslant \mathbf{q}^{(2)}, \quad \sigma \in D. \tag{7.57}$$

Find the controls $\boldsymbol{\lambda}(t)$ and $\mathbf{q}(t)$ subject to (7.10), (7.55) that minimize the total cost

$$V = \min_{\boldsymbol{\lambda}(t),\, \mathbf{q}(t)} \int_0^{\bar{t}} \left[\sum_{n=1}^{K} G_n(\lambda_n(t)) + \sum_{k=1}^{P} \Phi_k(q_k(t)) \right] dt, \tag{7.58}$$

where G_n (λ_n) and Φ_k (q_k) are the functions characterizing the cost of ejection reducing to λ_n, q_k, $n = \overline{1, K}$, $k = \overline{1, P}$, per unit time.

Problem 4. Suppose now that the economic resources used to reduce the ejections are bounded by some V_0, i.e.

$$\int_0^{t} \left[\sum_{n=1}^{K} G_n(\lambda_n(t)) + \sum_{k=1}^{P} \Phi_k(q_k(t)) \right] dt \leqslant V_0 \tag{7.59}$$

and that condition (7.56) does not hold for any $\boldsymbol{\lambda}(t)$, $\mathbf{q}(t)$ satisfying constraints (7.55), (7.56).

Under the above assumptions minimize the functional

$$J(\lambda, \mathbf{q}) = \int_0^{\bar{t}} \int_\sigma [\varphi(\mathbf{x}, t, \lambda, \mathbf{q}) - \tilde{\varphi}(\mathbf{x}, t)]^2 \, d\mathbf{x} \, dt, \tag{7.60}$$

where $\tilde{\varphi}(\mathbf{x}, t)$ is a given function characterizing permissible pollution in a region $\sigma \subset D$.

Neglecting condition (7.54) the minimum of the functional $J(\lambda, \mathbf{q})$ exists and problem (7.10), (7.55), (7.56), (7.60) allows in general an infinite set of solutions [7.32]. Consider in some detail the case $\mathbf{q}(t) \equiv 0$. Let $G_n(u)$, $\lambda_n^{(1)} \leqslant u \leqslant \lambda_n^{(2)}$, be convex functions. Then the set of functions $\lambda(t)$ that satisfies conditions (7.57), (7.59) is bounded and convex in $L_2(0, \bar{t})$ and consequently in problem (7.10), (7.55), (7.57), (7.59), (7.60) there are optimum controls [7.32], [7.33] which are characterized by the inequality

$$\sum_{n=1}^{K} \int_0^{\bar{t}} \varphi^*(\mathbf{x}_n, t, \mathbf{R}) [\rho_n(t) - \lambda_n(t)] \, dt \geqslant 0 \tag{7.61}$$

for every $\rho_n(t) \in L_2(0, \bar{t})$, $n = \overline{1, K}$, satisfying conditions of the form (7.57), (7.59). The function $\varphi^*(\mathbf{x}, t, \mathbf{R})$ is a solution of the corresponding adjoint problem (7.11) [7.8], [7.32], [7.38] for

$$\eta(\mathbf{x}, t) \equiv \varphi(\mathbf{x}, t, \mathbf{R}) - \tilde{\varphi}(\mathbf{x}, t). \tag{7.62}$$

To solve problem (7.10), (7.55), (7.57), (7.59), (7.60) we use the methods of minimizing quadratic functionals in Hilbert space.

All the problems considered in this paper are interrelated and the computing algorithms for their solution are constructed on a single constructive basis that is defined by the variational principle in conjunction with the splitting method. The study of the sensitivity of the models produces values of the gradients of functionals, generalized characteristics of models and simulated phenomena, that are required practically in all methods of optimization and optimum design. The software implementation of the models has a modular structure and is in the form of several application programs incorporated in the functional part of the software of the system of the automation of investigations in the physics of the atmosphere, ocean, and environmental protection.

References

7.1. G. I. Marchuk, *Mathematical Modelling in the Environmental Problem* (Moscow, Nauka, 1982) (in Russian).

7.2. G. I. Marchuk and V. V. Penenko, in: *Problems of Rational Nature Management and of the Control of the Quality of the Natural Environment*

of Northern Siberia (Yakutsk: Yakut. Filial SO AN SSSR, 1979): 28-37 (in Russian).
7.3. G. I. Marchuk, V. V. Penenko, A. E. Aloyan, and A. V. Protasov, in: *Urgent Problems of Applied Mathematics and Mathematical Modelling* (Novosibirsk: Nauka, 1982): 125-137 (in Russian).
7.4. G. I. Marchuk, *Numerical Solution of Problems in the Dynamics of the Atmosphere and Ocean* (Leningrad: Gidrometeoizdat, 1974) (in Russian).
7.5. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd ed. revised and enlarged (Moscow: Nauka, 1980) (in Russian).
7.6. A. A. Samarsky, *Introduction to the Theory of Difference Schemes* (Moscow: Nauka, 1971) (in Russian).
7.7. N. N. Yanenko, *The Fractional-Step Method of Solving Multidimensional Problems of Mathematical Physics* (Novosibirsk: Nauka, 1967) (in Russian).
7.8. V. V. Penenko, *Methods of Numerical Simulation of Atmospheric Processes* (Leningrad: Gidrometeoizdat, 1982) (in Russian).
7.9. M. E. Berlyand, *Current Problems of Atmospheric Diffusion and Atmospheric Pollution* (Leningrad: Gidrometeoizdat, 1975) (in Russian).
7.10. V. V. Penenko and N. N. Obraztsov, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 12, No. 11: 1207-1211 (1976) (in Russian).
7.11. J. P. Boris and D. L. Book, *J. Comp. Phys.* 11, No. 1: 38-69 (1973).
7.12. C. K. Forester, *J. Comp. Phys.* 23, No. 1, 1-22 (1977).
7.13. A. V. Panarin, *On the Construction of Two Monotonic Schemes of Increased Order of Accuracy in Space* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint, 1982) (in Russian).
7.14. G. I. Marchuk and V. V. Penenko, *Izv. Akad. Nauk SSSR, Ser. Fiz. Atmosf. i Okeana* 15, No. 11: 1123-1131 (1979) (in Russian).
7.15. A. V. Panarin, in: *Methods of Mathematical Modelling in Hydrodynamic Environmental Problems* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1983) (in Russian).
7.16. V. V. Penenko and N. N. Obraztsov, *Meteorologiya i Gidrologiya*, No. 1: 3-16 (1976) (in Russian).
7.17. A. V. Protasov and V. A. Gannochka, in: *Numerical Solution of Ocean Dynamics Problems* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1982): 24-34 (in Russian).
7.18. V. V. Penenko, ed., *Mathematical Methods of Experimental Design* (Novosibirsk: Nauka, 1981) (in Russian).
7.19. V. V. Nalimov, ed., *New Ideas in Experimental Design.* (Moscow: Nauka, 1969) (in Russian).
7.20. V. V. Penenko and V. F. Raputa, *Meteorologiya i Gidrologiya*, No. 8: 38-46 (1982) (in Russian).
7.21. V. V. Fedorov, *The Theory of Optimum Design* (Moscow: Nauka, 1971) (in Russian).
7.22. A. B. Uspensky and V. V. Fedorov. *Kibernetika*, No. 4 (1974) (in Russian).
7.23. V. V. Fedorov, *Kibernetika*, No. 1 (1975) (in Russian).
7.24. G. I. Marchuk, *Dokl. Akad. Nauk SSSR* 227, No. 5 (1976) (in Russian).
7.25. G. I. Marchuk, in: *Computational Methods in Mathematical Physics, Geophysics, and Optimal Control* (Novosibirsk: Nauka, 1978) (in Russian).
7.26. V. I. Gurman, *Singular Problems of Optimal Control* (Moscow: Nauka, 1977) (in Russian).
7.27. V. I. Gurman and A. I. Moskalenko, ed., *Optimal Control of Nature-Economy Systems* (Moscow: Nauka, 1980) (in Russian).
7.28. W. L. Gorr, S. A. Gustafson, and K. O. Kortanek, *Environment and Planning*, No. 4 (1972).
7.29. V. V. Penenko and V. F. Raputa, *Meteorologiya i Gidrologiya*, No. 3: 59-68 (1983) (in Russian).

7.30. O. F. Balatsky, *The Economics of Pure Air* (Kiev: Naukova Dumka, 1979) (in Russian).
7.31. F. P. Vasil'ev, *Numerical Methods of Solving Extremum Problems* (Moscow, Nauka, 1980) (in Russian).
7.32. J.-L. Lions, *Contrôle optimal de systèmes gouvernés pas des équations aux dérivées partielles* (Paris: Dunod, 1968).
7.33. F. P. Vasil'ev, *Lectures on the Methods of Solving Extremum Problems* (Moscow: Moscow Univ. Press, 1974) (in Russian).

8 Mathematical Modelling of Infectious Diseases

G. I. Marchuk, L. N. Belykh, and S. M. Zuev

The modern development of medicine and immunology is characterized by the penetration of mathematical methods and computing technology into these areas of knowledge. The interest of mathematicians in immunology is accounted for by the fact that immunology, along with genetics, is becoming one of the leading biological sciences. Characteristic of immunology is a high rate of accumulation of new facts, discoveries, and concepts, and their being quickly brought into clinical use. This has made it possible to consider immune processes as a complex dynamic system, and to use a systems approach to the study of immunologic effects. A systems approach to immunology is called for to "find the keys" to such important problems as immunotherapy, transplantation of tissues and organs, infectious diseases (especially their chronic forms), and cancer. The solution of these problems relies on an understanding of how the immune system works. Mathematical modelling is one of the more promising tools for gaining an insight into these laws.

This paper presents the results obtained, the current state of the art, and prospects of development of the indicated trend which these authors and their colleagues have been developing since 1975. Methodologically this study breaks down into two parts. The first part discusses questions of mathematical modelling of infectious diseases, and the second develops mathematical methods of processing laboratory and clinical data.

8.1. MODERN CONCEPT OF THE DEVELOPMENT OF ANTIVIRAL IMMUNE RESPONSE

In modern medical literature an "infectious disease" is understood to be the "representation of interrelations established between members of biocenosis one of which, owing to the pathogenicity mechanisms, is capable of existing in the other and the other, owing to the defense mechanism, can exert a counteraction to the pathogenic effect" [8.1]. It is known [8.2] that the principal way of an organism's defense against an infection is its immune system. The essence of an immune reaction to the invasion of a genotically heterologous material (an antigen) including contagia is

to produce specific material substrates (antibodies, killer cells) capable of neutralizing antigens.

Thus we shall think of an infectious virus disease as a conflict between a population of contagia (viruses) and an organism's immune system. In this connection we give a more detailed immunogenetic description of the antiviral immune response.

An immune reaction against organism-invading contagia of viral infections (influenza, measles, poliomyelitis, viral hepatitis, etc.) includes both types of immune response; a humoral reaction in which antibodies are produced by the B-lymphocytes system, and a cellular reaction in which cytotoxic T-lymphocyte-effectors accumulate. The cellular response is the main type of reaction which ensures the organism's defense since antibodies do not penetrate into the cells. The cytotoxic effector lymphocytes accumulated as a result of the immune response find and distroy virus-affected cells, acting as killers of their own organism's cells. Thus the antiviral immune response of the cellular type has as it were an autoimmune nature. But this is not a true autoimmune reaction. The latter means pathologic reactions of the immune system against its own normal intact cells or normal cellular antigenic substances. Killer lymphocytes, if there is an antiviral immunity, destroy *virus-affected cells* of their own organism. This appears to be the only way to clear the organism from viruses if, of course, the intracellular defense mechanisms (interferon, enzymes, controlling the replication of neucleic acids, etc.) of the virus-affected cells fail to deal with virus proliferation themselves.

The leading role played by the T-cellular immune response in antiviral immunity has been proved over the past decade. It appears that immunocompetent cells recognize the antigens of freely circulating viral particles, and the accumulated T-effectors armed with a receptor system against viral antigens find virus-affected cells, and, *acting against the viral particle*, kill the infected cells.

In 1974-1975 Zinkernagel [8.3] followed by other investigators proved that the processes of recognition, proliferating the effector T-lymphocytes and the cytotoxic effect of accumulated T-killers are triggered by a virus-transplantation antigen complex rather than by a viral antigen itself. So-called double recognition was discovered. The molecules of the transplantation antigens of the major complex of histocompatibility (H-2K and H-2D in mice and HLA-A, HLA-B, and HLA-C in man) act as primary molecules that form complexes with viruses. The term "double recognition" emphasizes that the recognizing receptor structures of lymphocytes are complementary to a virus-transplantation antigen complex rather than to a viral particle. To date it is not clear if a T-lymphocyte carries two closely situated receptors $\sqcup Y$, one of which signifies a transplantation antigen and the other signifies a viral one, or a single receptor that has

a generalized recognizing structure, ⊬ or ⊢ The latter variant (⊢) seems the most plausible.

The work of Zinkernagel was soon followed by studies (Benaceraff, Unanue [8.4]) that showed that the principle of double recognition is true not only for viral antigens but also for antigens of any other nature, including antigens that stimulate not a cellular but a humoral immune response, antibody genesis. Only in these cases do other antigens of the major complex of histocompatibility (Ia in mice and apparently HLA-D in man) act as primary molecules forming complexes with an antigen. The principal cells that bring antigens to helper T-lymphocytes to be recognized, antigen-presenting cells, are macrophages.

Thus the concepts of the events occurring when both viral and other antigens are recognized have been made more precise. It has turned out that T-lymphocytes recognize antigens only when in a complex with macrophagal antigens of the major complex of histocompatibility; the macrophage has turned out to be the first cell to interact with heterologous antigens.

It had been finally proved by the same time that the antigens of the major complex of histocompatibility (Ia, H-2K, and H-2D) play an outstanding role in cellular interactions. It is by them that macrophages and lymphocytes recognize one another. If these antigens are not identical on interacting cells, say have genetically determined differences, then no macrophage-T-lymphocyte, macrophage-B-lymphocyte, T-B, and T-T cooperations take place.

The discovery of double recognition and the importance of the identity of antigens in the major complex of histocompatibility for cooperation by the cells interacting in an immune response has resulted in a refined scheme of cellular events in the development of the response. It should be stressed that a refinement has taken place rather than a change. As in the earlier schemes (Petrov [8.2]) it is the macrophage, helper lymphocyte, and effector lymphocyte that interact. Helper lymphocytes are called Th (helper) when antibody genesis is triggered, and Ta (amplifier) when a cellular response is triggered. Effector lymphocytes are respectively labelled B and Te (effector). In previous schemes an antigen-T-helper receptor complex, which is presented to a macrophage, was taken to be the triggering signal. Now an antigen-macrophage transplantation antigen complex which is presented to a T-helper or T-amplifier, which results in their proliferation and in the appearance of the clones of the cells Ta and Th, that should be taken as the triggering signal. The same signal from a macrophage and a signal from the corresponding helper (Ta or Th) result in a stimulation of the effectors Te and B which gives rise to the clones of killers (effectors) and plasma cells, respectively. The killers (effectors) destroy virus-infected cells of their own organ-

ism and the plasma cells produce molecules of antibodies capable of neutralizing freely circulating viral particles. This is the immunogenetic foundation on which mathematical models of virus disease are constructed.

8.2. THE SIMPLEST MATHEMATICAL MODEL OF A VIRUS DISEASE

Simplifying as much as possible the above picture we can arrive at the following system of equations which we call the simplest model of a disease, and which was constructed by one of the authors in 1975 [8.5]:

$$\begin{aligned}
\dot{V} &= \frac{dV}{dt} = (\beta - \gamma F)\, V, \\
\dot{F} &= \frac{dF}{dt} = \rho C - \eta\gamma F V - \mu_f F, \\
\dot{C} &= \frac{dC}{dt} = \xi(m)\, \alpha F(t-\tau)\, V(t-\tau) - \mu_c (C - C^*), \\
\dot{m} &= \frac{dm}{dt} = \sigma V - \mu_m m.
\end{aligned} \tag{8.1}$$

In this model the process of a disease is described as follows. At time $t^0 = 0$ the organism is invaded by an initial dose of viruses V^0. This population begins to multiply and affect the cells of the target organ (m is the number of affected cells). In the blood, part of the viruses encounter F antibodies (receptors of immunocompetent cells), which leads to the stimulation of the immune system. After time τ, required for the immunocompetent cells to divide and differentiate, a numerous population (clone) of plasma cells forms (C is the number of plasma cells) whose main function is antibody production. Antibodies fix the viruses and the outcome of the disease depends on the struggle between them. If the viruses seriously affect the organ, there is a general aggravation of symptoms ($\xi(m) < 1$) and the operation of the immune system is impaired. The antibody production ρC is reduced and the chances of a recovery become worse. The quantities appearing in system (8.1) which are not explained above are positive constants.

For a detailed analytical and numerical study of the model see [8.6], [8.7]. Here we will restrict our discussion to the most interesting biologic interpretations of the properties of the solutions of the model.

It has turned out that the model reproduces the four known forms of a disease: subclinical, acute with a recovery, chronic, and a lethal outcome. The development of one form or another in a healthy organism infected with a small dose of viruses does not depend on

the infection dose, but is determined by the immunologic status of the organism with respect to a given type of viruses (by the set of model parameters).

It has been found that chronic forms of a disease are due to weak stimulation of the immune system and that a possible method for their cure is by aggravation of the disease. This discovery was used as a basis for constructing a model for such a type of cure [8.8] that helped to formulate a theory of aggravation or biostimulation.

The essence of the theory of biostimulation or aggravation is as follows. A new nonproliferating nonpathogenic biostimulator antigen is administered to an organism suffering from a stable chronic form of a disease increasing the dose in a certain time interval. The immune system begins to form a response against the new antigen. By administering large doses of biostimulators it is possible, by virtue of the competition between the two antigens (viruses and biostimulators) for the macrophages, to have the immune response against the viruses blocked, i.e. to have it react only to the biostimulators and "forget" about the contagia. This results in the concentration of viruses in the organism beginning to grow. After some time the biostimulators cease to be injected, and are removed rather quickly from the organism. The immune system alone remains to stand against the "chronic" antigen, but the situation has significantly changed. While the effect of the biostimulators lasted the virus concentration had considerably increased, and is now capable of inducing an effective immune response resulting in the complete removal of the viruses from the organism. Along with this it increases the immunologic barrier of the organism in respect to the given antigen, this due to memory cells, i.e. repeated contacts with the virus would lead to more favourable forms of the disease. It is clear that a considerable increase in the concentration of viruses in the organism may severely affect the organ. Biostimulation must therefore be combined with antipathogenic therapy that does not destroy the virus structure. Simulation has thus shown that artificial aggravation of the chronic disease by using biostimulators leads to complete self-recovery of the organism. Clinical application of the theory of aggravation to the treatment of pneumonia testifies in favour of this theory.

To conclude, the simplest model helped; to simulate the course of mixed infections [8.8], the temperature reaction of the organism when infected [8.9], and the treatment of chronic forms with the help of a stimulator of antibody production (SAP) [8.6], to elucidate the picture of a general morbidity from viral hepatitis [8.10]; to establish a linear relation of the mass of the affected organ and the observed biochemical index in hepatitis [8.10]; and to formulate some hypotheses about the appropriateness of a number of immunologic effects (switching of antibody synthesis, tolerance) [8.6], [8.7].

8.3. THE MODEL OF AN ANTIVIRAL IMMUNE RESPONSE

Based on the material of Sec. 8.1 the following variables of the model are introduced:

$V_f(t)$ is the number of "free" viruses (viral particles freely circulating in the organism and capable of multiplying in the cells of an organ sensitive to a given type of virus);

$M_V(t)$ is the number of stimulated macrophages (macrophages that have interacted with free viruses);

$H_E(t)$ is the number of helper T-lymphocytes involved in the cellular response;

$H_B(t)$ is the number of helper T-lymphocytes involved in the humoral response;

$E(t)$ is the number of effector T-cells (killers);

$B(t)$ is the number of immunocompetent B-lymphocytes capable of receiving the stimulation signal from the stimulated macrophages M_V and helpers H_B;

$P(t)$ is the number of plasma cells;

$F(t)$ is the number of antibodies;

$C_V(t)$ is the number of virus-affected cells of the organ; and

$m(t)$ is the nonfunctioning part of the virus-affected organ.

We make the following simplifying assumptions.

1. The number of macrophages M in the organism and that of cells C in the target organ are reckoned constant and sufficiently large for the increase in the stimulated macrophages M_V and affected cells C_V to be proportional to the number of free viruses V_f.

2. The reception of the stimulating signal by a lymphocyte leads to the formation of a clone of terminal cells after the time required for it to divide and proliferate. For the helpers to be stimulated requires a single signal from M_V, and for the cells E and B a double signal: from M_V and from the corresponding helper (H_E and H_B).

3. Part of the clone formed of terminal cells may again be aroused by the corresponding signal and stimulated to form a new clone. The remaining part performs other immune functions: help in stimulation, killer effect, and antibody production.

4. Helper lymphocytes H_E and H_B, having performed an act of interaction with lymphocytes E and B, respectively ("having rendered help" to E and B lymphocytes), terminate their life cycle.

5. For some time the infected cells C_V perform their normal functions. Their death is due either to the development of an irreversible viral damage or to their elimination by effectors E. Thus the affected mass of the organ is composed of the cells killed by the virus and those killed by effector lymphocytes. Therefore, a considerable damage of the organ results in an aggravation of the general

state of the organism and hence in a decrease in the effectiveness of the stimulation of its immune system.

These assumptions lead to a system of equations which we call a model of an antiviral immune response:

$$\begin{aligned}
\dot{V}_f &= nb_E C_V E + pb_m C_V - \gamma_m M V_f - \gamma_f V_f F - k\sigma C V_f,\\
\dot{M}_V &= \gamma_M M V_f - \alpha_M M_V,\\
\dot{H}_E &= b_H [\xi(m)\,\rho_H M_V H_E|_{t-\tau_H} - M_V H_E] - b_P M_V H_E E + \alpha_H (H_E^* - H_E),\\
\dot{H}_B &= b_H^{(B)} [\xi(m)\,\rho_H^{(B)} M_V H_B|_{t-\tau_H^{(B)}} - M_V H_B]\\
&\qquad - b_P^{(B)} M_V H_B B + \alpha_H^{(B)} (H_B^* - H_B),\\
\dot{E} &= b_P [\xi(m)\,\rho_E M_V H_E E|_{t-\tau_E} - M_V H_E E] - b_E C_V E + \alpha_E (E^* - E),\\
\dot{B} &= b_P^{(B)} [\xi(m)\,\rho_B M_V H_B B|_{t-\tau_B} - M_V H_B B] + \alpha_B (B^* - B),\\
\dot{P} &= b_P^{(P)} \xi(m)\,\rho_B M_V H_B B|_{t-\tau_B} + \alpha_P (P^* - P),\\
\dot{F} &= \rho_f P - \eta_f \gamma_f V_f F - \alpha_f F,\\
\dot{C}_V &= \sigma C V_f - b_E C_V E - b_m C_V,\\
\dot{m} &= \mu b_E C_V E + \eta b_m C_V - \lambda m.
\end{aligned} \tag{8.2}$$

Here all the coefficients are positive constants; the asterisk marks the values corresponding to normal constant levels in a healthy organism (when $V_f = 0$).

This model has been constructed comparatively recently [8.11] and has not yet been studied in detail. However, what has already been done shows that it embodies the basic regularities of the simplest model, that it can be used to study the various kinds of immunodeficiency, and that on the basis of the competition of T- and B-systems for the antigen there may arise conditions in the course of a disease that have not yet been described. All this will make it possible to acquire a deeper understanding of the mechanisms of pathologic processes. It should also be noted that a further extension of the model consists in allowing for local mechanisms for the implementation of the effect of killer cells and antibodies in a particular virus-affected organ [8.12].

8.4. MATHEMATICAL PROBLEMS OF MODELLING

All the models we have constructed satisfy the following basic criteria of being adequate to actual processes:

(i) global existence (i.e. for all $t \geqslant 0$) and uniqueness of a solution (ii) its nonnegativeness for all $t \geqslant 0$ under nonnegative initial conditions, and (iii) existence of conditions of the stability of a stationary solution corresponding to the state of the organism's health. The first criterion ensures that within the framework of the models no biologically absurd event should occur, namely, that the amount of some component of the process will attain an infinite value in a finite time. The second criterion guarantees the nonnegativeness of the model's components, as would be expected from the biological meaning of the described process. The third criterion guarantees the existence of an immunologic barrier V^* (the attraction zone of the state of a healthy organism) such that, when a healthy organism is infected with a small dose of viruses $V^0 < V^*$, the disease does not develop and for it to develop it is necessary that either the viruses should pass this barrier $V^0 > V^*$ or the stability conditions should be broken. If the third criterion is neglected, then within the framework of the models any contact with the virus should lead to the development of a disease, although in real life this far from being the case.

The proof of existence and uniqueness theorems and of a theorem on the nonnegativeness of our models is based on majorizing the models by a linear system and the continuity of solutions, respectively (see [8.6], [8.7], for example). As to stability, for models with a large number of variables it is convenient to cite the following result we are using, which for simplicity is formulated for a system of two equations without time lag and whose proof as a matter of fact is contained in [8.13].

Theorem 8.1. *Given a system*

$$\dot{v} = f(x, v), \quad \dot{x} = \varphi(x, v) \tag{8.3}$$

with functions f and φ ensuring the existence and uniqueness of its solution which allows the stationary solution $\{x, v\} = \{x^*, 0\}$, it follows from the exponential stability of this solution in the system

$$\dot{v} = f(x^*, v) = f_1(v), \quad \dot{x} = \varphi(x, 0) = \varphi_1(x) \tag{8.4}$$

that it is asymptotically stable in (8.3).

The breaking down of our models into two subsystems is easy to effect by imparting to them a biological meaning: our models describe a conflict between the vector of "defense forces of an organism" x and the vector of a "viral origin proper" v. It would be extremely useful to prove theorems of this kind so as to simplify the analysis of the existence, uniqueness, and nonnegativeness of solutions.

It seems that apart from the three indicated criteria of adequacy it is difficult, if not impossible, to carry out an analytic study of the models. The only efficient method of investigation therefore is

by numerical simulation on a computer, which creates the problem of constructing reliable computing methods of solution. The situation is aggravated by the fact that our models are systems of ordinary differential equations with lagging argument whose solutions' derivatives, as is known [8.14], may have discontinuities of the first order. It is therefore inappropriate to use methods for solving ordinary differential equations since this may lead to the loss of the order of approximation.

We have proved the following result [8.15]:

Theorem 8.2. *Given in* (a, b) *a piecewise continuous differentiable function* $f(t)$ *that possesses on differentiability intervals derivatives of up to the* l*th order inclusively* $(l > 1)$ *and has at points* $\{t_k\}$ *of* (a, b) $(k = \overline{1, N})$ *discontinuities of the first order (or discontinuities of the first order of the derivatives)* $\Delta f_k^{(i)}$ $(i = \overline{0, l};\ k = \overline{1, N})$, *there exists a unique function* $g(t)$, *computable from the formula*

$$g(t) = \sum_{k=1}^{N} \Theta(t - t_k) \sum_{i=0}^{l-1} \Delta f_k^{(i)} (t - t_k)^i / i!$$

and smoothing up to the $(l-1)$*st order, such that the function* $u(t) = f(t) - g(t)$ *is* $(l-1)$ *times continuously differentiable in* (a, b).

This theorem allows one to use a "smooth apparatus" in problems with an insufficient smoothness of its components. Using this theorem we have constructed a generalized Taylor series and interpolation polynomials for the solutions of our models $f(t)$ and modified the classical Runge-Kutta difference schemes [8.15].

8.5. MATHEMATICAL MODELS AND THE ANALYSIS OF EXPERIMENTAL DATA

In this section we consider two problems associated with the analysis of clinical and laboratory data and of the data of biological experiments. We shall be concerned with the estimation of the gravity of an organism's state from the values of the indices specific to the disease and with the problem of evaluating the coefficients of the models we have considered from observed trajectories of state variables.

8.5.1. Generalized index of the gravity of a diseased organism's state. Let us return to the mathematical model of the immune response (8.1) in which the last equation describes changes in the extent of an organism's affection. It is a certain abstract characteristic which in clinical practice must be assigned a particular indicator. In [8.10] results are reported of a comparison between the model and experimental data obtained for a group of hepatitis patients. It is shown that such an indicator is the biochemical index φ_b [8.16]. It is a scalar function of the vector X which characterizes the state

of a diseased organism. The vector components are various clinically observed biochemical indices. The coefficients of the function and its form were evaluated from the minimum condition of the mean square deviation at the disease height t_h of the values of the index from the estimate of the gravity of the state given by the physician.

Since, with $t = t_h$ fixed, the values of the biochemical index obtained for N patients are of a random nature and may be regarded as the realization of a random variable, for a comparison with the simulation results we obtained solutions corresponding to different sets of coefficients: $\alpha = \alpha_0 + \xi$, where α is the vector of coefficients, α_0 are values corresponding to the acute form of the disease, and ξ is the normally distributed random variable with zero expectation and standard deviation $k\alpha_0$. The coefficient of variation k was chosen so that the sampling distributions of $\varphi_b(t_h)$ and $m(t_h)$ fit in the best way. The experiment established that the best fit takes place for $k = 0.04$ and the experimental data do not contradict the hypothesis about the linear relation between $\varphi_b(t_h)$ and $m(t_h)$.

On the other hand, the analysis of the experimental data in [8.17] has shown that the average value of $\varphi_b(t)/\varphi_b(t_h)$ has an exponential dynamics. The same follows from model (8.1) for $m(t)$, $t \geqslant t_h$. Since $m(t)$ affects the nature of the immune response and hence the course of the disease and its outcome, the choice of therapy must take into account the value of $m(t)$ whose actual analogue is the biochemical index. But the biochemical index was constructed to estimate the state of hepatitis patients and cannot be used for other diseases. The need therefore arose to formalize the approach to the construction of this kind of indicators. In [8.7] one of the possible approaches is considered. This paper presents a more formal construction.

Let $s = 0, 1, 2, \ldots, r$ be an estimate of the gravity of a patient's state given by the clinicist at the height of the disease and $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ be the corresponding vector of the state. The function $\varphi(\mathbf{X})$ on which the functional $J = \| \varphi(\mathbf{X}) - s \|$ attains a minimum is called a generalized index of the gravity of a diseased organism's state. It is clear that the solution of the given problem is determined by the choice of the norm $\|\cdot\|$. We shall consider $\mathbf{X}$ as an n-dimensional continuous random variable with correlation matrix R and define $\varphi(\mathbf{X})$ as a predictor of a random variable s having a minimum mean square error. Then

$$J = \sum_{s=0}^{r} p_s \mathrm{E}\left[(\varphi(\mathbf{X}) - s)^2 | s\right],$$

where E is the expectation operator and p_s is the probability of state s. If $\mathrm{E}(\varphi(\mathbf{X}) \mid s) = s$ for all $s = 0, 1, \ldots, r$, then the given functional attains a minimum.

Let $\mathrm{E}(x_k \mid s) = s$ for all s and $k = 1, 2, \ldots, n$. Then it is possible to take as $\varphi(\mathbf{X})$ a linear combination of indices:

$$\varphi(\mathbf{X}) = \sum_{i=1}^{n} \alpha_i x_i, \quad \sum_{i=1}^{n} \alpha_i = 1,$$

since in this case $\mathrm{E}[\varphi(\mathbf{X}) \mid s] = s$. The coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ are evaluated from the minimum condition of the Lagrangian:

$$\tilde{J}(\alpha, \lambda) = \sum_s p_s \mathrm{E}\left[\left(\sum_{i=1}^{n} \alpha_i (x_i - s)\right)^2 \mid s\right] + \lambda\left(\sum_{i=1}^{n} \alpha_i - 1\right).$$

Differentiating $\tilde{J}(\alpha, \lambda)$ with respect to α_i, $i = \overline{1, n}$, and equating the derivatives to zero we obtain a system of algebraic equations in α and λ:

$$\sum_{i=1}^{n} \alpha_i \sum_s p_s \mathrm{E}[(x_k - s)(x_i - s) \mid s] + \lambda = 0, \quad k = \overline{1, n},$$

$$\sum \alpha_i = 1.$$

In the case where R is a diagonal matrix the solution of the given system is written as

$$\alpha_i = \left(\sum_{k=1}^{n} \prod_{j \neq k} R_{jj}\right)^{-1} \prod_{k \neq i} R_{kk}, \quad \lambda = -\left(\sum_{k=1}^{n} \prod_{j \neq k} R_{jj}\right)^{-1} \prod_{k=1}^{n} R_{kk}$$

and for the variance of the generalized index we get

$$\mathrm{E}(\varphi(\mathbf{X}) - s)^2 = \left(\sum_{k=1}^{n} 1/R_{kk}\right)^{-1}.$$

Since R_{kk}, $k = \overline{1, n}$, are positive and bounded, we have $\lim_{n \to \infty} \mathrm{E}(\varphi(\mathbf{X}) - s)^2 = 0$. Thus the introduction of a generalized index reduces the estimation error of the patient's state and makes it possible to go from the analysis of an n-dimensional variable to that of a scalar. Suppose now that the regressions $\mathrm{E}(x_k \mid s) = f_k(s)$, $k = \overline{1, n}$, are nonlinear. This case can be reduced to the one considered by transforming regressive relations to a linear form, for example, with the help of the polynomials $\eta_k(x_k)$, $k = \overline{1, n}$, so that

$$\mathrm{E}[\eta_k(x_k) \mid s] = s$$

Then the generalized index will be a linear combination of polynomials:

$$\varphi(\mathbf{X}) = \sum_{i=1}^{n} \alpha_i \eta_i(x_i).$$

If the vector $\mathbf{X}$ characterizes the state of some organ, then $\varphi_t(\mathbf{X}) = \varphi_t$, $t \geqslant t_h$, as a function of time represents the process of recovery of the functions of the given organ. It is naturally assumed that the form of relations $E(x_i \mid s)$ and the coefficients α_i, $i = \overline{1, n}$, remain unchanged in time. Otherwise one can use the approach of [8.7].

So the transition to the scalar characterization of the organism's state significantly simplifies the analysis of the dynamics of the process of the disease. It follows from model (8.1) that at the recovery stage ($t \geqslant t_h$), when the pathogenic effect of the antigen may be neglected, $m(t)$ and hence φ_t decrease exponentially (this is supported by the experimental results obtained in [8.17]), i.e.

$$\dot{\varphi}_t = -A\varphi_t, \quad \varphi_0 = 1, \quad A = \text{const} > 0, \quad t \geqslant t_h = 0. \tag{8.5}$$

This equation describes the recovery process of the functions by the organism in the average. However, individual patients' trajectories for this process have random deviations from this regularity. As already pointed out, good correspondence of model (8.1) to the experimental data for $t = t_h$ has resulted from small random changes of the coefficients. In [8.17], [8.18] therefore it was suggested for the description of the process of recovery of φ_t that model (8.5) should be supplemented by introducing in the coefficient A a Gaussian white noise of small intensity ε:

$$\dot{\varphi}_t = -(A + \sqrt{\varepsilon}\,\xi_t)\,\varphi_t, \quad \varphi_0 = 1, \quad 0 < \varepsilon \ll A.$$

Then the model of the process in question is the following stochastic differential equation in the Ito form [8.19]:

$$d\varphi_t = (0.5\varepsilon - A)\,\varphi_t\,dt + \sqrt{\varepsilon}\,\varphi_t\,dw_t,$$

where w_t is a Wiener process.

It has been shown that the last equation describes well the observed trajectories of the generalized index and that the properties of its solution allow one to use the method of maximum likelihood to estimate the coefficients A and ε from the experimental data [8.17].

The calculation of the estimates of the regeneration rate A from the observed trajectories in different groups of patients enabled different methods of treatment to be objectively compared. According to model (8.5) the method to which the greater value of A corresponds should be considered more effective.

The simple solution of the problem of evaluating the parameter A from observed trajectories has stimulated an extension of such an approach to the case of more complicated models. First of all it appears most natural to consider deterministic models as the result of averaging a real stochastic system [8.20]. Then the set of factors affecting the studied process but not included into the model can be

taken into account by introducing random disturbances. These questions are discussed in the next section.

8.5.2. The problem of evaluating the parameters of the model. Let the model be a system of differential equations:

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t, \alpha), \quad \mathbf{x}_0 = c = \text{const} > 0, \quad t \in [0, T], \quad (8.6)$$

where $x_t \in R^n$, $\alpha \in R^l$ and the vector function $\mathbf{f}(\mathbf{x}, \alpha)$ is linear in α. In what follows we shall require the existence of the first and second partial derivatives of the right-hand side with respect to $\mathbf{x}$. We therefore shall assume that $\mathbf{f}(\mathbf{x}, \alpha)$ satisfies that condition. The solution of problem (8.6) will be denoted by $\mathbf{x}_t(\alpha)$ assuming that $t \in [0, T]$. Here T is the transition time in the simulated system. In other words, in the interval $[0, T]$ there occurs the development of the immune response and the recovery of the organism's healthy state; $T \sim 40$ days for hepatitis, for example, and $T \sim 30$ days for pneumonia.

Suppose as a result of an experiment or clinical observations a trajectory of the state variables of the model is obtained. This means that there is a set $\Theta = \{t_0, t_1, \ldots, t_N\}$ and the values of the variables $\mathbf{X} = \{\mathbf{x}_{t_0}, \mathbf{x}_{t_1}, \ldots, \mathbf{x}_{t_N}\}$ observed at these moments. If the experiment uses a group of m animals, then we have a group of trajectories $\mathbf{X}_m = \{\mathbf{X}^i, i = 1, 2, \ldots, m\}$ in which $\{\mathbf{X}^i = \mathbf{x}_t^i, t \in \Theta\}$ is the observation on the ith animal. It may happen that the experiment will produce not trajectories but sets of the values of variables for $t \in \Theta$, i.e. $\mathbf{X}'_m = \{\mathbf{x}_t^i, i = 1, 2, \ldots, m_t, t \in \Theta\}$ and all the values $\mathbf{x}_t^i$ are statistically independent. This case occurs when it is possible to make only one observation on one animal. In any case it is assumed that the animals are of the same strain, so that the trajectories obtained may be regarded as resulting from the repetition of the experiment with a single organism. It is obvious that the actual trajectories have a random character and cannot therefore be implemented within the framework of model (8.6) for some $\alpha = \bar{\alpha}$.

Note at once that the random nature of the trajectories is due not only to the error in the observation of the state variables but also to the effects of the various factors not considered in model (8.6). Therefore, in describing actual trajectories, along with the observation error, it is necessary to consider the random fluctuations occurring in the studied system. The physical meaning of the fluctuations is as follows. The mathematical model (8.6) considers only those factors which determine the dynamics of the process and makes it possible to explain the mechanism of the effect being studied. In constructing the model many unessential relations are therefore discarded, and the process involved is thus "isolated" from the system in which it occurs. If our knowledge about the mechanism of the effect under study are true to fact, then those factors will be discarded which have no decisive impact on the process, and the model must describe

qualitatively its basic regularities. In model (8.6) these factors are uncontrollable, so in describing actual trajectories it is natural to consider their effects as random disturbances that cause short-term deviations of the trajectories from a solution $\mathbf{x}_t(\alpha)$.

The problem of evaluating the vector α from $\mathbf{X}_m$ or $\mathbf{X}'_m$ is that it is necessary to assign to the set of sampling trajectories $\mathbf{X}_m$ ($\mathbf{X}'_m$) one solution of the set of solutions $\{\mathbf{x}_t(\alpha), \alpha \in D \subset R^l\}$ of problem (8.6), i.e. it is required to construct a functional $J(\mathbf{x}_t(\alpha), \mathbf{X}_m)$ such that

$$\max_{\alpha \in D} J(\mathbf{x}_t(\alpha), \mathbf{X}_m) = J(\mathbf{x}_t(\alpha_m), \mathbf{X}_m),$$

where α_m is the estimate of the vector of coefficients.

It will be assumed that there is no observation error.

Suppose first that the observed trajectory $\bar{\mathbf{x}}_t$, $t \in \Theta$, is in the set of solutions of system (8.6), i.e. for some $\alpha = \bar{\alpha}$, $\mathbf{x}_t(\bar{\alpha}) = \bar{\mathbf{x}}_t$, $t \in \Theta$. Then the problem of evaluating $\bar{\alpha}$ may be solved by the methods of perturbation theory [8.21] according to the following scheme.

It is assumed that some (unperturbed) state of the system involved with the corresponding vector α_0 is known. It is also assumed that the observed state called perturbed is described by equation (8.6) for $\alpha = \alpha_0 + \delta$ and the deviations δ are small compared to α_0. On this basis the problem is linearized and δ and the parameters corresponding to the perturbed state are evaluated (for details see [8.21]).

In this case the perturbed state cannot be implemented within the framework of system (8.6) using the above method, for the observed trajectories are not in the set of solutions of the model (owing to the random nature of these trajectories). Let model (8.6) describe an unperturbed motion with a set of parameters α_0 and let the perturbed motion have corresponding to it $\alpha_t = \alpha_0 + \delta_t$, where δ_t, $t \in [0, T]$, is the function of time. Then every sampling trajectory $\mathbf{X}^i$ is associated with δ_t^i, $t \in [0, T]$. Since $\mathbf{X}^i$, $i = 1, 2, \ldots, m$, have a random nature, it is natural to consider δ_t^i, $i = 1, 2, \ldots, m$, as realizations of a random process $\{\delta_t, t \in [0, T]\}$, $\mathrm{E} \mid \delta_t \mid^2 < \infty$. Since we assume that the differences in the trajectories are due to purely random factors, we have $\mathrm{E}\delta_t = 0$ for all $t \in [0, T]$, and the presence of random fluctuations with respect to the trajectory $\mathbf{x}_t(\alpha_0)$ means that

$$(1/T) \int_0^T \delta_t \, dt \approx 0. \tag{8.7}$$

The exact equation holds in the limit as $T \to \infty$ if $\mathrm{cov}(\delta_t, \delta_{t+\tau}) = \mathrm{E}\delta_t\delta_{t+\tau} \underset{\tau \to \infty}{\longrightarrow} 0$. For a finite T this is true therefore if $\mathrm{cov}(\delta_t, \delta_{t+\tau})$ vanishes sufficiently fast, i.e. for $\tau \ll T$.

For correct representation of (8.7) we introduce a process ξ_t such that $\mathrm{E} \mid \xi_t \mid^2 < \infty$, $\mathrm{E}\xi_t = 0$, $t \in [0, T]$,

$$\operatorname{cov}(\xi_t, \xi_{t+\tau}) \xrightarrow[\tau \to \infty]{} 0, \quad \lim_{T \to \infty} (1/T) \int_0^T \int_0^T \mathrm{E}\xi_t^i \xi_s^j \, dt \, ds = g^{ij},$$

$$G = (g^{ij}), \quad i, j = 1, 2, \ldots, l.$$

We represent, $\delta_t = \xi_{t/\varepsilon}$, where $\varepsilon > 0$ is a small parameter, and say that for any $T > 0$, $q > 0$

$$\lim_{\varepsilon \to 0} P\left\{ \left| (1/T) \int_t^{t+T} \xi_{s/\varepsilon} \, ds \right| > q \right\} = 0.$$

is uniform in t. Such a representation may be explained as follows. Suppose, for example, that ξ_t and the state variables have the same characteristic time of change. Then the variable $\delta_t = \xi_{t/\varepsilon}$ is fast in comparison with the dynamics of $\mathbf{x}_t(\alpha_0)$ and we shall assume (8.7) to hold for sufficiently small values of ε. Thus the set of trajectories $\mathbf{X}_m$ will be considered to be a restriction to Θ of the set of realizations of a random process $\mathbf{x}_t$ which satisfies the system of equations

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t, \alpha_0 + \xi_{t/\varepsilon}), \quad \mathbf{x}_0 = c, \quad t \in [0, T]. \tag{8.8}$$

As shown in [8.22] for any $T > 0$, $\delta > 0$

$$\lim_{\varepsilon \to 0} P\{ \sup_{0 \leqslant t \leqslant T} |\mathbf{x}_t - \mathbf{x}_t(\alpha_0)| > \delta \} = 0.$$

Consequently, for small ε model (8.8) describes the process of random fluctuations of the trajectories with respect to the general regularity of $\mathbf{x}_t(\alpha_0)$, i.e. the actually observed pattern.

Then the problem of evaluating α_0 from the set $\mathbf{X}_m$ ($\mathbf{X}'_m$) consists in filtering the fast random variable $\xi_{t/\varepsilon}$. It will be assumed for simplicity that G is a diagonal matrix. Also note that by virtue of the linearity of $\mathbf{f}(\mathbf{x}, \alpha)$ in α a representation $\mathbf{f}(\mathbf{x}, \alpha) = B(\mathbf{x})\alpha$ is permissible, $B(\mathbf{x})$ being an $n \times l$ matrix whose elements are functions of the vector $\mathbf{x}$.

Then under the assumptions made earlier about the right-hand side of model (8.6) and the properties of the process ξ_t it can be shown following [8.22] that the normalized difference

$$\zeta_t^\varepsilon = (\mathbf{x}_t - \mathbf{x}_t(\alpha_0))/\sqrt{\varepsilon},$$

where $\mathbf{x}_t$ is a solution of (8.8), as $\varepsilon \to 0$ weakly converges in the interval $[0, T]$ to a Gaussian Markov random process ζ_t^0 which satisfies the system of equations

$$\dot{\zeta}_t^0 = A[\mathbf{x}_t(\alpha_0)]\zeta_t^0 + B[\mathbf{x}_t(\alpha_0)]G'\dot{\mathbf{w}}_t, \quad \zeta_0^0 = 0, \tag{8.9}$$

where $\mathbf{w}_t$ is a Wiener random process with a covariance matrix It, I is an $l \times l$ identity matrix and $A(\mathbf{x})$ is an $n \times n$ matrix:

$$A(\mathbf{x}) = (\partial f^j(\mathbf{x}, \alpha)/\partial x^i), \quad G' = (\sqrt{g^{ii}}).$$

Since the random variable in (8.8) is a fast one, ε may be assumed sufficiently small to approximately consider $\{\mathbf{x}_t - \mathbf{x}_t(\alpha_0), \ t \in [0, T]\}$ as a Gaussian Markov process satisfying the equation:

$$\frac{d}{dt}(\mathbf{x}_t - \mathbf{x}_t(\alpha_0)) = A[\mathbf{x}_t(\alpha_0)](\mathbf{x}_t - \mathbf{x}_t(\alpha_0)) + B[\mathbf{x}_t(\alpha_0)]\Gamma\dot{\mathbf{w}}_t, \tag{8.10}$$

where $\Gamma = \sqrt{\varepsilon}G'$. Since the form of the distribution of deviations $\mathbf{x}_t - \mathbf{x}_t(\alpha_0)$ is known, the vector α_0 can be chosen proceeding from the best fit of the theoretical and the observed distribution of trajectories, i.e. the functional $J(\mathbf{x}_t(\alpha), \mathbf{X}_m)$ can be constructed by the principle of maximum likelihood:

$$J(\mathbf{x}_t(\alpha), \mathbf{X}_m) = \prod_{i=1}^{m} P\{\mathbf{x}_t(\alpha), \mathbf{X}^i\}.$$

Since the process is a Markov one,

$$P\{\mathbf{x}_t(\alpha), \mathbf{X}^i\} = \prod_{i=1}^{N} p(\mathbf{x}_{t_i} - \mathbf{x}_{t_i}(\alpha) | \mathbf{x}_{t_{i-1}} - \mathbf{x}_{t_{i-1}}(\alpha)),$$

where $p(\mathbf{x} \mid \mathbf{y})$ is the distribution density of $\mathbf{x}$ for a fixed value of $\mathbf{y}$. By (8.10) this density is Gaussian:

$$p(\mathbf{x}_{t_i} - \mathbf{x}_{t_i}(\alpha) \mid \mathbf{x}_{t_{i-1}} - \mathbf{x}_{t_{i-1}}(\alpha))$$
$$= [(2\pi)^n \det K_{t_i}]^{-1/2} \exp\{-(1/2)(\mathbf{x}_{t_i} - \mathbf{x}_{t_i}(\alpha) - \mathbf{M}_{t_i})^T$$
$$\times K_{t_i}^{-1}(\mathbf{x}_{t_i} - \mathbf{x}_{t_i}(\alpha) - \mathbf{M}_{t_i})\}, \tag{8.11}$$

where K_t and $\mathbf{M}_t$ are defined by the following equations in $[t_{i-1}, t_i]$:

$$\dot{\mathbf{M}}_t = A[\mathbf{x}_t(\alpha)]\mathbf{M}_t, \ \mathbf{M}_{t_{i-1}} = \mathbf{x}_{t_{i-1}} - \mathbf{x}_{t_{i-1}}(\alpha),$$

$$\dot{K}_t = A[\mathbf{x}_t(\alpha)]K_t + K_t A^T[\mathbf{x}_t(\alpha)] + F[\mathbf{x}_t(\alpha), \Gamma'], \tag{8.12}$$

$$K_{t_{i-1}} = 0, \quad F^{ij} = b^i \Gamma' (b^j)^T, \quad \Gamma' = \varepsilon G,$$

b^i is the ith row of the matrix $B[\mathbf{x}_t(\alpha_0)]$.

We shall consider the function

$$L(\alpha, \Gamma') = (1/m) \ln J(\mathbf{x}_t(\alpha), \mathbf{X}_m);$$

then

$$L(\alpha, \Gamma') = (1/m) \sum_{i=1}^{m} \sum_{j=1}^{N} \ln p[\mathbf{x}^i_{t_j} - \mathbf{x}_{t_j}(\alpha) | \mathbf{x}_{t_{j-1}} - \mathbf{x}^i_{t_{j-1}}(\alpha)]. \tag{8.13}$$

Given a set of independent values $\mathbf{X}'_m = \{\mathbf{x}_t^i,\ t \in \Theta,\ i = \overline{1, m}\}$, rather than the set of trajectories $\mathbf{X}_m$, there is

$$L(\alpha, \Gamma') = \sum_{j=1}^{N} (1/m_j) \sum_{i=1}^{m_j} \ln p_i[\mathbf{x}_{t_j}^i - \mathbf{x}_{t_j}(\alpha)] \qquad (8.14)$$

and equations (8.12) are solved in the interval $[0, T]$ with initial conditions $\mathbf{M}_0 = 0$, $K_0 = 0$ since $\mathbf{x}_0 - \mathbf{x}_0(\alpha) = 0$. It follows from (8.12) that $\mathbf{M}_t = 0$, $t \in [0, T]$. If in this case we assume K_t to be a diagonal matrix with time-independent elements, then (8.14) goes over into a sum of squared deviation of $\mathbf{x}_t(\alpha)$ from $\mathbf{X}'_m$ which corresponds to the least-squares technique. This case occurs when the random character of the trajectories is due entirely to observation error. But if the observation error is to be considered together with the perturbations of a real system, then the matrix K_t in (8.11) should be replaced by a matrix $K'_t = K_t + S$, where S is the diagonal matrix of error variance.

The analysis of the function of likelihood has shown that if for all $\alpha \in D$ solutions of equations (8.12) and (8.6) exist and there is in addition a set Θ such that for $\alpha_1 \neq \alpha_2$, $\alpha_1, \alpha_2 \in D$, one can always find $t_k \in \Theta$ for which $\mathbf{x}_{t_k}(\alpha_1) \neq \mathbf{x}_{t_k}(\alpha_2)$, then the limiting point of the sequence α_m, $m = 1, 2, \ldots$, will with probability 1 be α_0, where α_0 is an unknown true value of the vector of parameters. The same results can also be obtained for systems nonlinear in α.

A similar reasoning can be used also in the case where the right-hand side of the model contains variables with time lag:

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-\tau}, \alpha),\ \mathbf{x}_s = \varphi(s),\ s \in [-\tau, 0]. \qquad (8.15)$$

The stochastic model for which (8.15) is an averaged system is the equation

$$\dot{\mathbf{x}}_t = \mathbf{f}(\mathbf{x}_t, \mathbf{x}_{t-\tau}, \alpha + \xi_{t/\varepsilon}).$$

It appears that by a reasoning similar to that of [8.22] it can be shown for the system with time lag that $(\mathbf{x}_t - \mathbf{x}_t(\alpha))/\sqrt{\varepsilon}$ will be a Gaussian process, and this will allow the method of maximum likelihood to be used to estimate α from $\mathbf{X}_m$ [8.23].

References

8.1. O.V. Baroyan and D. R. Porter, *International and National Aspects of Modern Epidemiology and Microbiology* (Moscow: Meditsina, 1975) (in Russian).

8.2. R. V. Petrov, *Immunology and Immunogenetics* (Moscow: Meditsina, 1976) (in Russian).

8.3. R. M. Zinkernagel, *J. Exp. Med.*, No. 143: 437-443 (1976).

8.4. B. Benaceraff and E. R. Unanue, *Textbook on Immunology* (Baltimore: Williams and Wilkins, 1979).
8.5. G. I. Marchuk, *The Simplest Mathematical Model of a Virus Disease* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, Preprint, 1975) (in Russian).
8.6. L. N. Belykh and G. I. Marchuk, in: *Mathematical Modelling in Immunology and Medicine* (Novosibirsk: Nauka, 1982): 5-27 (in Russian).
8.7. G. I. Marchuk, *Mathematical Methods in Immunology* (New York: Optimization Software Inc., Publ. Division, 1983).
8.8. L. N. Belykh, in: *Mathematical Modelling in Immunology and Medicine* (Novosibirsk: Nauka, 1982): 33-40 (in Russian).
8.9. A. L. Asachenkov, *ibid.*, pp. 40-44.
8.10. A. A. Romanyukha, *ibid.*, pp. 27-33.
8.11. G. I. Marchuk and R. V. Petrov, *A Mathematical Model of an Antiviral Immune Response* (Moscow: Otdel Vychisl. Matem. AN SSSR, Preprint No. 40, 1983) (in Russian).
8.12. G. I. Marchuk and R. V. Petrov, *Virus Affection of an Organ and Immunophysiological Defense Reactions (A Mathematical Model)* (Moscow: Otdel Vychisl. Matem. AN SSSR, Preprint No. 40, 1983) (in Russian).
8.13. N. N. Krasovsky, *Some Problems of the Theory of the Stability of Motion* (Moscow: Fizmatgiz, 1959) (in Russian).
8.14. L. E. Elsgolts and S. B. Norkin, *Introduction to the Theory of Differential Equations with Deflective Argument*, 2nd ed., rev. and enl. (Moscow: Nauka, 1971) (in Russian).
8.15. L. N. Belykh, in: *Proc. of the IFIP-TC-7 Working Conf., Moscow, USSR, 5-11 July, 1982* (Amsterdam: North-Holland, 1983): 79-84.
8.16. G. I. Marchuk, N. I. Nisevich, and I. I. Zubikova, in: *Mathematical Methods in Clinical Practice* (Novosibirsk: Nauka, 1978): 27-33 (in Russian).
8.17. I. B. Pogozhev, I. I. Zubikova, and A. A. Romanyukha, *ibid.*, pp. 40-55.
8.18. S. M. Zuev, in: *Mathematical Modelling in Immunology and Medicine* (Novosibirsk: Nauka, 1982): 93-100 (in Russian).
8.19. K. J. Aström, *Introduction to Stochastic Control Theory* (New York: Academic Press, 1970).
8.20. S. M. Zuev, in: *Proc. of the IFIP-TC-7 Working Conf., Moscow, USSR, 5-11 July, 1982* (Amsterdam: North-Holland, 1983): 255-264.
8.21. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd ed., rev. and enl. (Moscow: Nauka, 1980) (in Russian).
8.22. A. D. Ventsel and M. I. Freidlin, *Fluctuations in Dynamic Systems under Small Random Perturbations* (Moscow: Nauka, 1979) (in Russian).
8.23. L. N. Belykh, S. M. Zuev, and G. I. Marchuk, in: *Proc. of the IFIP-TC-7 Conf. on Recent Advances in System Modeling and Optimization, Hanoi, 1983* (Berlin, Heidelberg, New York: Springer Verlag, 1983).

9 Software for Designing Electrophysical Devices

V. P. Ilyin, A. V. Gavrilin, B. I. Golubtsov, N. I. Gorbenko, V. A. Kateshov, G. S. Popova, V. M. Sveshnikov, A. L. Urvantsev, and M. V. Urev

A system for automatically designing (SAPR) (for Sistema Avtomatizirovannogo Proektirovaniya) electrophysical devices which can calculate the characteristics of a construction and so enable its parameters to be optimized is considered. We discuss the calculation of two- and three-dimensional electro- and magnetostatic fields in ungeometrical regions, wave processes in dynamic systems, and the simulation of high-current beams of charged particles. In mathematical terms these topics can be reduced to boundary value and Cauchy problems for elliptic equations, to partial eigenvalue problems for boundary value problems, to the integral equations of potential theory, and to systems of the differential equations which describe the motion of charged particles given a space charge.

We shall give mathematical statements and the basic algorithms for these problems and cover their implementation in the form of software packages.

9.1. INTRODUCTION

Traditionally electrophysics has been concerned with the calculation of electromagnetic fields, current densities, and charge distribution. It is applied in many fields of science and technology and in the design of high-voltage constructions, electrical devices, transformers, charged particle accelerators, spectrum analysers, image converter tubes, microwave electronic devices, high-current beam generators, semiconductor devices, and plasma handles. The number and variety of electrophysical devices makes it necessary for us to be able to study them both in terms of simulating constructions that already exist and designing new systems with predetermined properties.

9.2. PHYSICAL MODELS AND MATHEMATICAL STATEMENTS

In the case of stationary potential electro- and magnetostatic fields the problem of calculating them can be reduced to solving Poisson's equation in a region Ω with boundary Γ (see [9.1])

$$\Delta u \equiv \frac{1}{x^\alpha}\frac{\partial}{\partial x}\left(x^\alpha \frac{\partial u}{\partial x}\right) + \frac{\partial^2 u}{\partial y^2} + \beta \frac{\partial^2 u}{\partial z^2} = f(x, y, z), \quad (9.1)$$

where $\beta = 1$, $\alpha = 0$ for three-dimensional problems, $\beta = 0$, $\alpha = 1$ for three-dimensional axisymmetric problems, and $\beta = 0$, $\alpha = 0$ for plane-parallel problems.

On the boundary Γ or parts of it Γ_i ($\cup\Gamma_i = \Gamma$), $i = 1, 2, 3$, one of the following conditions may be given:

$$u|_{\Gamma_1} = g_1(x, y, z), \tag{9.2}$$

$$\partial u/\partial \mathbf{n} \mid_{\Gamma_2} = g_2(x, y, z), \tag{9.3}$$

$$[\partial u/\partial \mathbf{n} + \gamma(x, y, z)\, u]\,|_{\Gamma_3} = g_3(x, y, z). \tag{9.4}$$

In addition, if there are media with different physical properties, the so-called conjugation condition

$$u|_{\Gamma_4^+} = u|_{\Gamma_4^-}, \quad \varepsilon^+ \partial u/\partial \mathbf{n}|_{\Gamma_4^+} = \varepsilon^- \partial u/\partial \mathbf{n}|_{\Gamma_4^-} \tag{9.5}$$

is imposed at their interface Γ_4. The solution of problem (9.1)-(9.5) is assumed to be unique and to have sufficient smoothness, unless otherwise stated. Besides the interior boundary value problem (9.1)-(9.5) we shall consider an equivalent system of integral equations for the surface density of charges $\mu(Q)$, $Q \in \Gamma$, which in the three-dimensional case is of the form (see [9.2])

$$\begin{gathered}
\int_\Gamma \mu(Q)\, G_1(P, Q)\, d\Gamma = g_1(P), \quad P \in \Gamma_1, \\
2\pi\mu(P) + \int_\Gamma \mu(Q)\, [\partial G_1(P, Q)/\partial \mathbf{n} + \gamma(Q)\, G_1(P, Q)]\, d\Gamma \\
= g_3(P), \quad P \in \Gamma_3, \\
2\pi\mu(P) + [(\varepsilon_+ - \varepsilon_-)/(\varepsilon_+ + \varepsilon_-)] \int_\Gamma \mu(Q)\, [\partial G_1(P, Q)/\partial \mathbf{n}]\, d\Gamma = 0, \quad P \in \Gamma_4,
\end{gathered} \tag{9.6}$$

where the kernel $G_1(P, Q) = 1/R(P, Q)$. From the known density $\mu(Q)$ it is possible to find the potential

$$u(P) = \int_\Gamma [\mu(Q)/R(P, Q)]\, d\Gamma. \tag{9.7}$$

To get the exterior boundary value problem given by the second and third equations of system (9.6) it is necessary to change signs before the integrated terms of the left-hand sides.

The calculation of a two-dimensional nonlinear magnetostatic field without hysteresis can be reduced to finding the z-component (θ-component in the axisymmetric case) of the magnetic potential vector A as the solution of the boundary value problem

$$-\frac{\partial}{\partial x_1}\left[\nu_r(|\mathbf{B}|^2)\left(\frac{\partial A}{\partial x_1} + \alpha\frac{A}{x_1}\right)\right] - \frac{\partial}{\partial x_2}\left[\nu_r(|\mathbf{B}|^2)\frac{\partial A}{\partial x_2}\right] = \mu_0 j \text{ in } \Omega, \tag{9.8}$$

where $\alpha = 0$, $x_1 = x$, $x_2 = y$ in the plane case and $\alpha = 1$, $x_1 = r$, $x_2 = z$ in the axisymmetric case, $\mathbf{B} = (B_{x_1}, B_{x_2})$, $B_{x_2} = -\partial A/\partial x_2$, $B_{x_1} = \partial A/\partial x_1 + \alpha A/x_1$, $\mathbf{j}$ is the current density vector, μ_0 is the vacuum permeability and the nonlinear function of the magnetization ν_r ($|\mathbf{B}|^2$) satisfies the following relations:

$$\nu_r(x) \in C^2[0, \infty), \quad \nu_r'(x) \geqslant 0, \quad x \in [0, \infty). \tag{9.9}$$

The boundary conditions (9.2) to (9.4) may be given on the boundary Γ or its parts Γ_i. Instead of the boundary value problem (9.8), (9.9), (9.2)-(9.4) consider the functional of magnetic field energy (see [9.3])

$$F(A) = \int_{\Omega} \left[\int_0^{|\mathbf{B}|^2} \nu_r(\xi)\, d\xi - \mu_0 A j \right] d\Omega \tag{9.10}$$

for which equation (9.8) is the Euler equation. We shall assume that the solution A is in a corresponding Sobolev space. We have described the statements of the problems of calculating two- and three-dimensional electrostatic and magnetostatic fields.

Now we will describe briefly the class of electrodynamics problems (see [9.4]). Consider an electromagnetic field in a vacuum ($\varepsilon = \mu = 1$). It is convenient to introduce the scalar and vector potentials, φ and $\mathbf{A}$:

$$\mathbf{H} = \operatorname{curl} \mathbf{A}, \quad \mathbf{E} = -\operatorname{grad} \varphi - (1/c)\, \partial \mathbf{A}/\partial t, \tag{9.11}$$

which satisfy the wave equations

$$\Delta \mathbf{A} - (1/c^2)\, \partial^2 \mathbf{A}/\partial t^2 = -4\pi \mathbf{j}/c, \quad \Delta \varphi - (1/c^2)\, \partial^2 \varphi/\partial t^2 = -4\pi\rho \tag{9.12}$$

and the relation (the Lorentz calibration condition):

$$\operatorname{div} \mathbf{A} + (1/c)\, \partial \varphi/\partial t = 0. \tag{9.13}$$

In a number of cases it is more convenient to represent $\mathbf{E}$ and $\mathbf{A}$ as sums of potential and curl components

$$\mathbf{E} = \mathbf{E}^{(p)} + \mathbf{E}^{(c)}, \quad \mathbf{A} = \mathbf{A}^{(p)} + \mathbf{A}^{(c)} \tag{9.14}$$

satisfying the equations

$$\begin{gathered} \operatorname{div} \mathbf{A}^{(c)} = \operatorname{div} \mathbf{E}^{(c)} = \operatorname{curl} \mathbf{E}^{(p)} = 0, \\ \mathbf{E}^{(c)} = -(1/c)\, \partial \mathbf{A}/\partial t, \quad \mathbf{E}^{(p)} = -\operatorname{grad} \varphi, \quad \mathbf{H} = \operatorname{curl} \mathbf{A}^{(c)}. \end{gathered} \tag{9.15}$$

The new scalar and vector potentials φ and $\mathbf{A}^{(c)}$, as expected from (9.14)-(9.15), satisfy the system of equations

$$\begin{gathered} \Delta \mathbf{A}^{(c)} - (1/c^2) \partial^2 \mathbf{A}^{(c)}/\partial t^2 = -4\pi \mathbf{j}/c - (1/c)\, \partial\, (\operatorname{grad} \varphi)/\partial t, \\ \Delta \varphi = -4\pi\rho. \end{gathered} \tag{9.16}$$

A specific feature of this statement is that the scalar potential φ is determined by the Coulomb interaction of charges and only para-

metrically depends on time. The vector potential may be represented as the series

$$A^{(c)} = \sum_{k=1}^{\infty} q_k(t) A_k. \tag{9.17}$$

where A_k are the eigenfunctions of the operator $\Delta = \text{grad div} - \text{curl curl}$ defined by the relations

$$\Delta A_k + \lambda_k A_k = 0, \quad \text{div } A_k = 0, \quad A_k^{(t)}|_{\Gamma} = 0. \tag{9.18}$$

The last condition means that the tangential component of the vector A_k is zero on the metallic conductor surface. Problem (9.18) defines a discrete set of eigenvalues λ_k. The eigenfunctions corresponding to distinct λ_k are mutually orthogonal and may be assumed to be normalized according to the condition

$$\int_{\Omega} A_k A_l \, dV = \delta_{kl}$$

In the problem of exciting oscillations of resonant cavities λ_k and A_k are the frequencies and harmonics of free oscillations. To determine the amplitudes $q_k(t)$ the first equation of (9.16) is multiplied by A_k and integrated with respect to the volume of the cavity Ω. Considering relations (9.17)-(9.18) we obtain the equation

$$\ddot{q}_k + \omega_k^2 q_k = (4\pi/c) \int_{\Omega} j A_k \, dV, \quad \omega_k = \lambda_k c \tag{9.19}$$

After finding the vector potential the electric and magnetic intensities are evaluated from the equations

$$\mathbf{E}^{(c)} = -(1/c)\sum_k \dot{q}_k A_k, \quad \mathbf{H} = \sum_k q_k \, \text{curl } A_k. \tag{9.20}$$

Thus the main computational problem is the partial eigenvalue problem (9.18). To approximate the vector potential by a finite segment of series (9.17) it is required to find several of the smallest eigenvalues λ_k and the corresponding eigenfunctions A_k.

In simulating high-current beams the mathematical statement consists in solving the so-called self-consistent nonlinear problems which can be reduced to calculations of the electromagnetic fields and the motion of beams of charged particles taking into account space charges and currents (see [9.5]).

The potential of the electric field φ formed by a system of electrodes and the space charge of particles with density ρ satisfies Poisson's equation

$$\Delta\varphi = -4\pi\rho. \tag{9.21}$$

The motion of particles with a mass M and a charge z is given by the equation (in Gauss units)

$$d\,(M\,d\mathbf{r}/dt)/dt = z\mathbf{E} + (z/c)\,[(d\mathbf{r}/dt)\,\mathbf{H}], \tag{9.22}$$

where $\mathbf{r}$ is the radius vector of a particle, c is the velocity of light, t is the time, and $\mathbf{H}$ is the magnetic vector.

The flow continuity condition of charges

$$\operatorname{div}\mathbf{j} = \operatorname{div}\rho\mathbf{v} = \partial\rho/\partial t \tag{9.23}$$

is satisfied, with $\mathbf{j}$ the current density and $\mathbf{v}$ the average velocity defined as

$$\mathbf{v}\,(r) = \lim_{\Delta V\to 0}\,[\sum z_i\mathbf{v}_i\,(r)/\sum z_i]. \tag{9.24}$$

Here $\mathbf{v}_i = d\mathbf{r}_i/dt$ is the velocity of an individual particle with a charge z_i. The sums are taken over all particles in volume ΔV. Accordingly the density of the space charge is of the form

$$\rho = \lim_{\Delta V\to 0}\,[\sum z_i/\Delta V] = q\,(\Delta V)/\Delta V, \tag{9.25}$$

where $q\,(\Delta V) = \sum z_i$ is the total charge in ΔV.

The magnetic field is represented as the sum $\mathbf{H} = \mathbf{H}_e + \mathbf{H}_p$, where $\mathbf{H}_e$ is the intensity of the external field which is given or calculated as a function of independent variables. The vector $\mathbf{H}_p$ specifies the magnetic field proper of the beam

$$\mathbf{H}_p(r) = (1/c)\int_G \{[\mathbf{j}\,(\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')]/|\mathbf{r}-\mathbf{r}'|^3\}\,dV. \tag{9.26}$$

It is required to find a solution, i.e. the distribution of the potential, the current density, space charges, and the trajectories of particles in a closed region G on whose boundary Γ the conditions for φ and $\mathbf{j}$, and the initial data for the coordinates and velocities of the particles are given.

The problem of optimizing axisymmetric electron-optical systems (EOS) reduces to finding an axial potential distribution such that the value of a functional F, which characterizes the optical properties of EOS, reaches a minimum.

To solve the problem the potential is represented as

$$\Phi\,(z) = \Phi_0(z) + \sum_k \Phi_k(z)\,\delta\beta_k, \tag{9.27}$$

where $\Phi_0\,(z)$ is the axial distribution for the initial boundary problem (the initial estimate of the boundary conditions and geometry of the region), $\Phi_k\,(z)$, the disturbance influence functions, are the

derivatives $\partial\Phi/\partial\beta_k$ of the axial potential with respect to the parameters defining the disturbances of the geometry of the boundary or boundary conditions, and $\delta\beta_k$ is the value of the given (kth) disturbance.

Thus optimization reduces to finding disturbances $\delta\beta_k$ such that F reaches a minimum.

The functional to be minimized is represented as $F = \sum_{i=1}^{N} a_i (J_i - J_{0,i})^2$, where $N = 1, 2, 3$, a_i, $J_{0,i}$ are the given numbers, and J_i are the characteristics of the electron-optical image which are the functions of the axial distribution $\Phi(z)$.

To describe the process of calculating the geometric influence function a parameter β_k which specifies the variation of the boundary is given. The following variations are provided for:

(a) a translation of a part of the boundary along the direction unit vector $\mathbf{e} = (e_r, e_z)$, with β_k the value of that translation;

(b) a longitudinal extension (compression) of a rectilinear segment of the boundary along the direction vector (q_r, q_z) with a fixed endpoint, with β_k the length of extended segment;

(c) a rotation of a part of the boundary about a fixed point, with β_k the angle of rotation;

(d) a change in the radius of the spherical generator; one endpoint of the generator and its radius are on the z-axis and the other endpoint changes either along the z-axis (r = const) or along the r-axis (z = const), with β_k the radius of the generator.

It is usual to require in such problems that there should occur no discontinuities when there is a disturbance. In such cases one variation involves several portions of curves and a forced variation not coinciding in type with the main, governing, variation is given on them.

The influence function for geometric disturbances is given by the formula

$$\Phi_k(z) = \int_\Gamma (\sigma\psi_k + \sigma_k' G)\, d\tau, \tag{9.28}$$

where $\psi_k = \partial G/\partial\beta_k$, $\sigma_k' = \partial\sigma/\partial\beta_k$, and σ and G are the density and the kernel of the original integral equation.

The density disturbance function is given by the integral equation

$$\int_\Gamma \sigma_k'(P)\, G(P, Q)\, dS_Q = \int_\Gamma \sigma(Q)\, [\varphi_k(P, Q) + \psi_k(P, Q)]\, dS_Q. \tag{9.29}$$

9.3. NUMERICAL METHODS OF SOLVING ELECTROPHYSICAL PROBLEMS

As is seen from the preceding section, there is a great variety in the statements of electrophysical problems and it is therefore difficult to choose a universal approach to their numerical solution which would take into account both the specific features of simulated physical processes and the specific character of the mathematical statements themselves.

To solve the potential problems (9.1)-(9.15) and electrodynamic problems ([9.1], [9.4], [9.5]) we consider the use of the finite-difference method (FDM).

Consider a grid formed by coordinate lines (or planes) parallel to the axes and covering a region G: $x = x_i$, $y = y_j$, $z = z_k$, $i = 0, 1, \ldots, l$, $j = 0, 1, \ldots, m$, $k = 0, 1, \ldots, n$. Here and in what follows there is no third coordinate for the two-dimensional case.

We denote by (i, j, k) a node with coordinates (x_i, y_j, z_k) and let $h^{(1)}_{i,j,k} = h^x_{i-1} = x_i - x_{i-1}$, $h^{(2)}_{i,j,k} = h^y_{j-1} = y_j - y_{j-1}$, $h^{(3)}_{i,j,k} = h^x_i = x_{i+1} - x_i$, $h^{(4)}_{i,j,k} = h^y_j = y_{j+1} - y_j$, $h^{(5)}_{i,j,k} = h^z_{k-1} = z_k - z_{k-1}$, $h^{(6)}_{i,j,k} = h^z_k = z_{k+1} - z_k$ denote the distances from (i, j, k) to the adjacent nodes. We introduce the notation $h = \max\limits_{i,j,k} \{\sup h^{(r)}_{i,j,k}\}$. The node (i, j, k) will be called internal with respect to G if $(x_i, y_j, z_k) \in G$ and external otherwise. We denote by Ω_0 the set of all internal nodes, by Ω_1 the set of external nodes, and by Ω_Γ the set of boundary points of the grid, i.e. the points of the intersection of coordinate lines of the grid with the boundary Γ of G.

Let $u = \{u(x_i, y_j, z_k)\} = \{u_{i,j,k}\}$ be the grid function defined at the grid nodes. Then its range forms an N-dimensional vector, where N is the total number of the nodes involved. If the index of the node (i, j, k) ordered in some way is L, then $u_{i,j,k} = u_L$. For example, when the natural numbering is used $L = i + jl + klm$, $i = 0, 1, \ldots, l$; $j = 0, 1, \ldots, m$; $k = 0, 1, \ldots, n$.

The differential equations defining the boundary-value problem will be written in operator form

$$L\mathbf{v} = \mathbf{g}. \tag{9.30}$$

To the differential problem (9.30) is assigned a system of algebraic finite-difference equations

$$L_h\mathbf{u} = \mathbf{f}, \tag{9.31}$$

where $\mathbf{u}$ and $\mathbf{f}$ are N-dimensional vectors whose components are de-

fined as the values of the grid function at the nodes of the grid Ω and L_h is a square matrix.

On denoting grid nodes by Q_k system (9.31) may be written as

$$P_0(Q_k)\,\mathbf{u}(Q_k) - \sum_{Q_i \in D(Q_k)} P_i(Q_k)\,\mathbf{u}(Q_i) = \mathbf{f}(Q_k), \qquad (9.32)$$

where $D(Q_k)$ is the collection of nodes used in the equation for a point Q_k (these nodes will be called adjacent to Q_k) and $s(Q_k)$ is the number of such nodes.

The matrix of the system of difference equations will be symmetric subject to the condition $P_i(Q_k) = P_{i'}(Q_{k'})$ if the nodes k and i', k' and i coincide pairwise.

When a sufficiently large number of nodes are used, Poisson's equation can be approximated up to any given error with the aid of finite-difference expressions. We give simple but most frequently occurring examples. The simplest "one-dimensional" difference Laplacian is of the form (the indices j, k are omitted for brevity)

$$(\Lambda_x u)_i = \frac{2u_{i-1}}{h^x_{i-1}(h^x_{i-1}+h^x_i)} - \frac{2u_i}{h^x_{i-1}h^x_i} + \frac{2u_{i+1}}{h^x_i(h^x_{i-1}+h^x_i)}$$

$$= \left(\frac{\partial^2 u}{\partial x^2}\right)_i + \frac{h^x_i - h^x_{i-1}}{3}\left(\frac{\partial^3 u}{\partial x^3}\right)_i$$

$$+ \frac{(h^x_i)^2 - h^x_i h^x_{i-1} + (h^x_{i-1})^2}{12}\left(\frac{\partial^4 u}{\partial x^4}\right)_i O(h^3) \qquad (9.33)$$

In particular, for a uniform grid ($h^x_i = h_x$) we have

$$(\Lambda_x u)_i = \frac{1}{h_x^2}(u_{i-1} - 2u_i + u_{i+1})$$

$$= \left(\frac{\partial^2 u}{\partial x^2}\right)_i + \frac{h_x^2}{12}\left(\frac{\partial^4 u}{\partial x^4}\right)_i + \frac{h_x^4}{360}\left(\frac{\partial^6 u}{\partial x^6}\right)_i + O(h^6). \qquad (9.34)$$

The approximation error of these expressions on the functions having bounded derivatives up to the third order inclusively is $O(h)$. But if the grid is uniform and the function has a fourth derivative, then the error has a second order.

The simplest approximations for Poisson's equation are:

$$(\Delta_5 u)_{i,j} = ((\Lambda_x + \Lambda_y)\,u)_{i,j} = f_{i,j} \qquad (9.35a)$$

for the plane problem and

$$(\Delta_7 u)_{i,j,k} = ((\Lambda_x + \Lambda_y + \Lambda_z)\,u)_{i,j,k} = f_{i,j,k} \qquad (9.35b)$$

for three variables, with Λ_y and Λ_z defined similarly to Λ_x in (9.33).

Obviously both difference equations (9.35a), (9.35b) approximate Poisson's equation with an error of the first order on a nonuniform

grid and an error of the second order on a uniform one. More precisely, $|\psi| \leqslant M_3 h/3$ for the former case and $|\psi| \leqslant M_4 h^2/12$ for the latter, with

$$M_k = \max_l \sup_{S_{i,j}} \partial^k u / \partial x^l \, \partial y^{k-l},$$

$$l = 1, \ldots, k, \quad S_{i,j} = \{x_{i-1} \leqslant x \leqslant x_{i+1}, \quad y_{j-1} \leqslant y \leqslant y_{j+1}\}.$$

The above finite-difference expressions approximating the differential equations are defined if the points used are in the set of nodes Ω_0 internal with respect to G. For the nodes near the boundary, however, the difference equations may use points lying on the boundary Γ of G or outside G. In that case when a differential equation is approximated we assume that its solution is smoothly extended beyond the region for a distance h. The external grid nodes that appear in the difference equations are sometimes called fictitious. The internal nodes whose difference equations contain points on the boundary of or outside the region will be called near-boundary nodes and the remaining internal nodes will be called regular nodes.

For a system of difference equations to be definite, i.e. for the number of equations to equal that of the grid nodes appearing in the system, it is necessary to use finite-difference approximations of the boundary conditions. The methods of approximating boundary conditions may differ depending on the character of location of the boundary nodes with respect to the boundary, the grid patterns, and the order of the approximation error. The problems differ considerably when the boundary conditions contain derivatives and when they do not.

For the basic algorithms and the specific features of the approximation of boundary conditions, as well as other schemes of the approximation of differential operators, see [9.1], [9.6], for example.

The obtained systems of algebraic equations can be solved by direct, iterative and mixed methods, including the multigrid methods and iterations over subregions without overlapping (see [9.1], [9.8], [9.9]).

The general form of the iterative processes for solving the system of equations $A\mathbf{u} = \mathbf{f}$ ($\mathbf{u}$, $\mathbf{f}$ are vectors and A is a square matrix of the Nth order) is given by the formula

$$\mathbf{u}^n = \mathbf{u}^{n-1} + H_n (\mathbf{f} - A\mathbf{u}^{n-1}), \quad n = 1, 2, \ldots, \tag{9.36}$$

where $\mathbf{u}^n$ is the vector of the nth approximation (the nth iteration) and H_n are nonsingular matrices specifying particular iterative algorithms). Relation (9.36) may be written in an equivalent form

$$\mathbf{u}^n = B_n \mathbf{u}^{n-1} + \mathbf{g}^n, \tag{9.37}$$

where $\mathbf{g}^n = H_n \mathbf{f}$ and $B_n = E - H_n A$ is the transition matrix or the iteration step matrix.

A large number of direct iterative methods and ways of optimizing them are given in [9.7].

Methods of solving systems of algebraic equations based on combining direct and iterative methods have lately been actively developed. The idea behind these methods is that the computational region may be partitioned into subregions so that one or more of them are rectangular and include the major part of the grid nodes. Then the solution of the original problem can be reduced to a successive solution of auxiliary problems in the subregions so that the iterative approximations converge in the region as a whole to the desired solution. It is necessary to give special boundary conditions on the boundaries of the introduced subregions for each approximation.

It is essential in this case to construct well converging iterations in the subregions.

Suppose, for example, that the computational region G is partitioned into two subregions $G^{(1)}$ and $G^{(2)}$ by an auxiliary curve S_1 and that there is an approximation to the desired solution in $G^{(2)}$ and on the boundary between the subregions, S_1: $\mathbf{u}_{2k}^{(2)}(x, y)$, $(x, y) \in G^{(2)} \cup S_1$. Then we find a solution $\mathbf{u}_{2k+1}^{(1)}$ of the problem in $G^{(1)}$ given boundary conditions on S_1, whose right-hand side depends on $\mathbf{u}_{2k}^{(2)}$. Then in $G^{(2)}$ we find the next approximation $\mathbf{u}_{2k+1}^{(2)}$ given the boundary conditions on S_1 defined taking into account $\mathbf{u}_{2k+1}^{(1)}$, the process then being repeated iteratively:

$$\begin{aligned}
&\alpha_1 \mathbf{u}_{2k+1}^{(1)} + \beta_1 \partial \mathbf{u}_{2k+1}^{(1)}/\partial \mathbf{n}|_{S_1^-} = \alpha_1 \mathbf{u}_{2k}^{(2)} + \beta_1 \partial \mathbf{u}_{2k}^{(2)}/\partial \mathbf{n}|_{S_1^+}, \\
&\alpha_2 \mathbf{u}_{k+2}^{(2)} + \beta_2 \partial u_{2k+2}^{(2)}/\partial \mathbf{n}|_{S_1^+} = \alpha_2 \mathbf{u}_{2k+1}^{(1)} + \beta_2 \partial \mathbf{u}_{2k+1}^{(1)}/\partial \mathbf{n}|_{S_1^-}, \\
&\qquad\qquad k = 0, 1, 2, \ldots .
\end{aligned} \tag{9.38}$$

Here $v|_{S_1^-}$ and $v|_{S_1^+}$ are limiting values of the functions as they tend to S_1 from $G^{(1)}$ and $G^{(2)}$, respectively. Each of the approximations $\mathbf{u}_{2k+1}^{(1)}$ and $\mathbf{u}_{2k+2}^{(2)}$ is defined as a solution of Poisson's equation in $G^{(1)}$ and $G^{(2)}$ with the original boundary conditions on the boundary of the region G or conditions (9.38) on S_1. The iterative process can be modified by introducing relaxation parameters:

$$\begin{aligned}
&\Delta \hat{\mathbf{u}}_{2k+1}^{(1)} = f^{(1)}, \quad (x, y) \in G^{(1)}, \\
&\mathbf{u}_{2k+1}^{(1)} = \omega_1 \hat{\mathbf{u}}_{2k+1}^{(1)} + (1 - \omega_1)\, \mathbf{u}_{2k-1}^{(1)}, \\
&\Delta \hat{\mathbf{u}}_{2k+2}^{(2)} = f^{(2)}, \quad (x, y) \in G^{(2)}, \\
&\mathbf{u}_{2k+2}^{(2)} = \omega_2 \hat{\mathbf{u}}_{2k+2}^{(2)} + (1 - \omega_2)\, \mathbf{u}_{2k}^{(2)},
\end{aligned} \tag{9.39}$$

with $\mathbf{u}^{(1)}_{2k+1}$ and $\mathbf{u}^{(2)}_{2k+2}$ on the left-hand sides of (9.38) replaced by $\hat{\mathbf{u}}^{(1)}_{2k+1}$ and $\hat{\mathbf{u}}^{(2)}_{2k+2}$. Without going into the studies of algorithms of this kind we note the principal result of the latest work: there are values α_i, β_i, ω_i, $i = 1, 2$, such that the iterative process (9.38), (9.39) converges to the solution of the original problem, i.e. $\mathbf{u}^{(1)}_{2k+2} \to \mathbf{u}(x, y)$ for $(x, y) \in G^{(1)}$ and $\mathbf{u}^{(2)}_{2k+2} \to \mathbf{u}(x, y)$ for $(x, y) \in G^{(2)}$ as $k \to \infty$. One variant of the boundary conditions is as follows: a Dirichlet condition is given on S_1 for $G^{(1)}$ and a Neumann condition is given for $G^{(2)}$, i.e. $\beta_1 = \alpha_2 = 0$. In the general case the problem of optimizing α_1, β_1, α_2, β_2, ω_1, ω_2 remains open (they may be variable as well over different iterations).

Schwartz's alternating method and iterations over nonoverlapping subregions extend in a natural way to the case of partitioning the original region into several subregions and to three-dimensional problems.

We have described the iterations over the subregions at the differential level, but when difference equations are solved, expressions (9.38), (9.39) are naturally replaced by the corresponding finite-difference approximations. This approach combined with the efficient direct methods of the fast Fourier transform or cyclic reduction when an ungeometrical region is partitioned into simple subregions appears to be promising since it may be expected that the rate of convergence of the iterations over the subregions is independent of the grid steps.

Another hybrid way of solving the boundary-value problems by grid methods is by the multigrid method.

One of the particular versions of such an approach for solving the equation $\Delta_n \mathbf{u}_n = \mathbf{f}_n$ consists in a successive performance of several procedures:

(a) The iterative process

$$\mathbf{u}^n_h = \mathbf{u}^{n-1}_h + \alpha\,(\Delta_h \mathbf{u}^{n-1}_h - \mathbf{f}_n), \quad n = 1, 2, \ldots, N_1, \tag{9.40}$$

with α and N_1 chosen so that the computed components of the $\mathbf{u}^{N_1}_h$ expansion into a Fourier series in the matrix eigenvectors are sufficiently small; for example, for the Dirichlet problem in a square the eigenvalues Δ_h are $\lambda_{p,q} = -(4/h^2)\,[\sin^2(ph/2) + \sin^2(qh/2)]$, $h = \pi/(N+1)$, $p, q = 1, 2, \ldots, N$. Assuming those harmonics to be high-frequency for which p or q are greater than $N/2$, we shall have an error suppression coefficient, with $\alpha = h^2/5$, less than $q \approx 0.6$. In $N_1 = 13$ iterations it is thus possible to suppress these components by a factor of more than 10^3.

(b) The residual of the approximation is obtained

$$\mathbf{r}_n = \mathbf{f}_n - \Delta_n \mathbf{u}^{N_1}_n \tag{9.41}$$

and a vector $\mathbf{r}_{n_1}$ made up of the components equal to the values of $\mathbf{r}_n$ coincident nodes of grids Ω_n and Ω_{n_1} is determined.

(c) The equation

$$\Delta_{n_1}\mathbf{z}_{n_1} = \mathbf{r}_{n_1} \tag{9.42}$$

is solved on the grid Ω_{n_1}

(d) The vector

$$\mathbf{z}_n = p_{2n}\mathbf{z}_{n_1} \tag{9.43}$$

is determined using the operator of interpolating $P_{2n,n}$ from Ω_{n_1} to Ω_n.

If the values of $\mathbf{z}_{i,j}$ on a square grid are only known for even i and j, then a simple interpolation is carried out using the formulas

$$\mathbf{z}_{i,j} = \begin{cases} (\mathbf{z}_{i+1,j} + \mathbf{z}_{i-1,j})/2, & \text{with } i \text{ odd and } j \text{ even,} \\ (\mathbf{z}_{i,j+1} + \mathbf{z}_{i,j-1})/2, & \text{with } i \text{ even and } j \text{ odd.} \\ (\mathbf{z}_{i-1,j} + \mathbf{z}_{i,j+1} + \mathbf{z}_{i,j-1} + \mathbf{z}_{i+1,j})/4, & \text{with } i \text{ and } j \text{ odd.} \end{cases} \tag{9.44}$$

(e) The approximation on Ω_n is corrected

$$\mathbf{u}_n^{N_1+1} = \mathbf{u}_n^{N_1} + \mathbf{z}_n. \tag{9.45}$$

Next N_2 iterations are carried out using formula (9.40) and $\mathbf{u}_n^{N_1+1}$ as the initial estimate ($N_2 \leqslant N_1$). The residual $\mathbf{f}_n - \Delta_n \mathbf{u}_n^{N_1+N_2+1}$ is again calculated and, if it is still large, the entire cycle is repeated. As a rule, the approximation $\mathbf{u}^{N_1+N_2+1}$ now coincides, within the required accuracy with the desired difference solution $\mathbf{u}_n$.

For the correction of the system of equations (9.42) may in turn be solved using a still sparser grid Ω_{h_2} and so on. If the solution obtained on a sparse grid Ω_{h_k} is used to construct the initial estimate for $\Omega_{h_{k-1}}$, then it can be shown (by sufficiently complicated manipulations we omit) that the total volume of calculations required for the solution $\mathbf{u}_h$ to be obtained is $O(h^{-2})$, i.e. is proportional to the number of nodes.

When approximating integral equations in the two-dimensional case individual parts of the boundary are described in parametric form and the potential density is represented by the function of the parameter $\sigma = \sigma(t)$ which is approximated on each simple connected part of the boundary using B-splines. If the values of $\sigma(t)$ are known at $N + 1$ points t_k of the interval $[a, b]$, then the B-spline approximation is given by the formula

$$\sigma_n(t) = \sum_{k=-p}^{N+p} c_k B_k^{(n)}(t), \tag{9.46}$$

where $B_k^{(n)}(t)$ is a B-spline of the nth order,

$$B_k^{(n)}(t) = (n+1) \sum_{l=k-\alpha}^{N+p} (\bar{t}_l - t)_+^n / \omega'_{k-l}(\bar{t}_l), \quad k = 0, 1, \ldots, \tag{9.47}$$

$$\omega_k(t) = (t - \bar{t}_{k-\alpha})(t - \bar{t}_{k+1-\alpha}) \ldots (t - \bar{t}_{k-\beta}).$$

If $(n + 1)$ is an odd number $\alpha = n/2$, $\beta = \alpha + 1$, $p = \alpha$, otherwise $\alpha = \beta = (n + 1)/2$, $p = (n - 1)/2$. In (9.47) the $\bar{t}_k$ are the spline nodes which for $n = 3$ coincide with t_k and for $n = 2$ are half-way between t_{k-1} and t_k.

To uniquely define the spline interpolant at the ends of the interval a, b boundary conditions for one of the following forms are given:

$$\sigma'_n(a) = A, \quad \sigma''_n(a) = C, \quad \sigma'_n(b) = B, \quad \sigma''_n(b) = D, \tag{9.48}$$

where A, B, C, and D are numbers chosen a priori by considering the behaviour of the interpolated functions at the ends of the interval. If $\sigma(t)$ is periodic at a part of Γ (for example, Γ is a closed electrode and the points a and b coincide), then periodicity conditions are imposed on the spline: $\sigma'_n(a) = \sigma'_n(b)$, $\sigma''_n(a) = \sigma''_n(b)$.

Note that the part of the boundary on which one spline $\sigma_n(t)$ is constructed may in fact consist of segments of different type, i.e. be a collection of different segments of straight lines or circles, on which different boundary conditions for the potential may even be given.

An approximate solution of the integral equations is constructed by the collocation principle as follows. Let the boundary of a region be made up of L simple-connected segments Γ_l, $l = 1, 2, \ldots, L$, whose equations are represented in parametric form. Each of the intervals $[\alpha_l, \beta_l]$ of the change of the parameter τ is partitioned uniformly into N_l subintervals by points τ_k, $k = l - 1 + \sum_{l'=1}^{l-1} N'_l, \ldots$ $\ldots, l + \sum_{l'=1}^{l} N'_l$. The total number of points is thus $N = L + \sum_{l=1}^{L} N_l$. For each of them an approximation of one of the integral relations is written, depending on the type of the boundary condition on the corresponding part of the boundary, with the function $\sigma(\tau)$ simultaneously replaced by its spline interpolant. We thus obtain N algebraic equations in $2pL + N$ unknown coefficients c_k. Adding then $2pL$ equations from the boundary conditions for the splines we obtain a system of equations of the $(2pL + N)$th order

$$A\mathbf{c} = \mathbf{f} \tag{9.49}$$

with a square matrix $A = \{a_{i,j}\}$ and a vector of unknowns $\mathbf{c} = \{c_i\}$. The elements of the matrix A are obtained from the rows

corresponding to the boundary conditions for the spline by a straightforward substitution of (9.46) to (9.48) and are easily found. The determination of the elements of the remaining rows requires calculating auxiliary integrals.

The numerical solution of integral equations for three-dimensional problems, while not differing in principle from the case of two-dimensional problems, is much more complicated technically. The challenge is naturally the increased dimensionality. It is necessary to approximate the density function on two-dimensional manifolds and to carry out the numerical integration with respect to surfaces that often have a complicated configuration in practical problems. The specific features of the surface potential density now occur in the neighbourhoods of dihedral or trihedral angles and its asymptotic behaviour has a more complicated form. If the characteristic discretization step for the boundary is h, then the order of the resulting system of linear algebraic equations is $O(h^{-2})$. To ensure a given accuracy we must solve dense systems with a thousand or more equations, which requires large computer resources.

As noted above, the solution of the problem of calculating the magnetostatic field in a two-dimensional region Ω reduces to finding the function $A(\mathbf{x})$, $\mathbf{x} = (x_1, x_2) \in \Omega$ minimizing the energy functional (9.10). Then the problem of seeking a minimum of (9.10) is replaced by an approximation problem. To do this Ω is approximated by a region Ω_h made up of closed triangles T_j which form a grid ω_h and have only one vertex or one side in common. A minimum of (9.10) is sought on Ω_h in the class V_h of continuous functions linear on every T_j. The space V_h is a finite-dimensional one and one can choose Courant functions φ_i as a basis, where i is the index of a node of ω_h. Then for any function $v \in V_h$ we can write the expansion

$$v(x) = \sum_{i=1}^{N} v_i \varphi_i(x), \tag{9.50}$$

where N is the total number of nodes in ω_h and v_i is the corresponding constant equal to the value v_i at the ith node. The function $A^h \in V_h$ minimizing the approximate problem will satisfy the system of nonlinear algebraic equations:

$$\partial F(A_h)/\partial A_i = 0, \quad i = 1, 2, \ldots, N. \tag{9.51}$$

One of the efficient ways of solving such a system is the Newton method and its various modifications, the one-step SOR-Newton method [9.3], for example. This last is as follows. Some initial estimate A_h^0 is chosen and to find the value of the $(n+1)$st iteration at the ith node the formula

$$A_i^{n+1} = A_i^n + \omega \frac{\partial F(A_1^{n+1}, A_2^{n+1}, \ldots, A_{i-1}^{n+1}, A_i^n, \ldots, A_N^n)/\partial A_i}{\partial F(A_1^{n+1}, A_2^{n+1}, \ldots, A_{i-1}^{n+1}, A_i^n, \ldots, A_N^n)/\partial A_i^2} \tag{9.52}$$

is used with $\omega \in (0, 2)$. The process is iterated until adjacent iterations are close in some norm. For the first and second derivatives of the functional we have

$$\frac{\partial F}{\partial A_i} = (2\pi)^{\alpha} \int_{\Omega} \left\{ \frac{\nu_r}{\mu_0} \left(B_{x_1} \frac{\partial B_{x_1}}{\partial A_i} + B_{x_2} \frac{\partial B_{x_2}}{\partial A_i} - j \frac{\partial A}{\partial A_i} \right\} x_1^{\alpha} \, d\Omega, \quad (9.53)$$

$$\frac{\partial^2 F}{\partial A_i^2} = (2\pi)^{\alpha} \int_{\Omega} \left\{ \frac{\partial (|\mathbf{B}|^2)}{\partial A_i} \nu_r' \frac{1}{\mu_0} \left(B_{x_1} \frac{\partial B_{x_1}}{\partial A_i} + B_{x_2} \frac{\partial B_{x_2}}{\partial A_i} \right) \right.$$

$$+ \frac{\nu_r}{\mu_0} \left[\left(\frac{\partial B_{x_1}}{\partial A_i} \right)^2 + \left(\frac{\partial B_{x_2}}{\partial A_i} \right)^2 + B_{x_1} \frac{\partial^2 B_{x_1}}{\partial A_i^2} \right.$$

$$\left. \left. + B_{x_2} \frac{\partial^2 B_{x_2}}{\partial A_i^2} \right] - j \frac{\partial^2 A}{\partial A_i^2} \right\} x_1^{\alpha} \, d\Omega. \quad (9.54)$$

Suppose that for an arbitrary triangle T_s of a triangulation ω_h the nodes are consecutively numbered i, j, k. Then, setting

$$a_i = x_{1,j} x_{2,k} - x_{1,k} x_{2,j}, \quad b_i = x_{2,j} - x_{2,k}, \quad c_i = x_{1,k} - x_{1,j},$$

$$a_j = x_{1,k} x_{2,i} - x_{1,i} x_{2,k}, \quad b_j = x_{2,k} - x_{2,i}, \quad c_j = x_{1,i} - x_{1,k}, \quad (9.55)$$

$$a_k = x_{1,i} x_{2,j} - x_{1,j} x_{2,i}, \quad b_k = x_{2,i} - x_{2,j}, \quad c_k = x_{1,j} - x_{1,i},$$

for the values of A^h inside the triangle T_s we get

$$A^h(x_1, x_2) = \frac{(a_i + b_i x_1 + c_i x_2) A_i^h + (a_j + b_j x_1 + c_j x_2) A_j^h + (a_k + b_k x_1 + c_k x_2) A_k^h}{2\Delta_s}, \quad (9.56)$$

where Δ_s is the area of T_s. Using (9.56) we get for the expressions in (9.53) and (9.54)

$$B_{x_1} = (-1)^{\alpha} \frac{c_i A_i^h + c_j A_j^h + c_k A_k^h}{2\Delta_s}, \quad (9.57)$$

$$B_{x_2} = (-1)^{\alpha+1} \left[\frac{b_i A_i^h + b_j A_j^h + b_k A_k^h}{2\Delta_s} + \alpha \left(\frac{A_i^h + A_j^h + A_k^h}{x_{1,i} + x_{1,j} + x_{1,k}} \right) \right], \quad (9.58)$$

$$\partial B_{x_1} / \partial A_i = (-1)^{\alpha} c_i / 2\Delta_s,$$

$$\frac{\partial B_{x_2}}{\partial A_i} = (-1)^{\alpha+1} \left[\frac{b_i}{2\Delta_s} + \alpha \left(\frac{1}{x_{1,i} + x_{1,j} + x_{1,k}} \right) \right], \quad (9.59)$$

$$\frac{\partial^2 B_{x_1}}{\partial A_i^2} = 0, \quad \frac{\partial^2 B_{x_2}}{\partial A_i^2} = 0, \quad \frac{\partial A^h}{\partial A_i} = \frac{a_i + b_j x_1 + c_i x_2}{2\Delta_s}, \quad \frac{\partial^2 A^h}{\partial A_i^2} = 0.$$

By representing the integrals in (9.53) and (9.54) as the sums of integrals taken over the triangles which form the triangulation ω_h, then substituting expressions (9.57) to (9.59) into the integrals, and approximating the functions $\nu_r(|\mathbf{B}|^2)$ and $\nu_r'(|\mathbf{B}|^2)$ inside the triangle by their values at the centre of T_s we obtain a final expression to calculate A_i^{n+1}. Also efficient, besides method (9.52), has proved to be the method of iterating over nonoverlapping subregions. The computational region may be partitioned into subregions Ω_1 with a nonlinear function ν_r and into Ω_2 with a linear function. To obtain a solution in Ω_1 we can use method (9.52), for example, and to obtain one in Ω_2 any iterative or direct method of solving a linear problem.

In solving the eigenvalue problems which arise in the study of problems of electrodynamics different numerical algorithms, such as the finite-difference method, are efficiently used. To find a minimum eigenvalue one can use the power method with a variable spectrum shift and to find several eigenvalues one can employ the method of iterations in a subspace (see [9.10], [9.11]).

In simulating strong current beams for the calculation of space charges and current density we use the large particle model. For stationary problems a more economical stream-tube model is used.

The physical character of the problems under consideration is greatly dependent on the type of boundary and initial conditions.

The calculation is carried out in a closed region on individual sections of whose boundary either Dirichlet or Neumann conditions are given. If media with different dielectric constants ε_+ and ε_- are present, then the conjugation condition is given at the interface.

As a rule we are interested in regions with ungeometrical configurations, including multiply connected ones. It is assumed that there are desired beams of particles of various sort in the region. The current density is given or the condition of the space-charge limitation of the current density is imposed.

Self-consistent stationary problems are solved by successive approximations with respect to the space charge. First a field is found for some initial distribution ρ (as a rule a zero one) and the trajectories and currents and the charge due to them are calculated in such a field. Then another field is found by taking into account the charge, the trajectories are again calculated and so on. The iterations continue until the adjacent approximations coincide to a given accuracy. A solution of a nonstationary problem is defined to be a sequence of stationary solutions for each of which the boundary and initial conditions correspond to time $t_n = \sum_{k=1}^{n} \Delta t_k$. The values of Δt_k are chosen sufficiently small compared to the characteristic time of change in the parameters of the problem.

Boundary-value problems for potentials are solved using difference methods on rectangular grids with piecewise constant steps in each of the independent variables. To calculate the intensity of the electrostatic field, algorithms for the first or second order approximation are used.

With an external magnetic field, it is assumed that its components are given at the points of the symmetry axis. A spline interpolation is made using these values and then the intensity at the interior points of the region is determined from a Scherzer series. When choosing an algorithm for calculating the trajectories, we are first interested in algorithms which are relatively easily implementable for solving practical problems since the trajectories are calculated in regions with ungeometrical configurations. Second, the electric field gradients are calculated numerically from the potentials at the nodes of the difference grid. In this case it requires a large number of operations to calculate the electric and magnetic components and to analyze the positions of the points of the trajectories.

For these reasons we prefer methods that have a small number of calculations on the right-hand side for each integration step. We use economical modifications for different types of problems depending on the presence of a magnetic field and the allowance for relativism. Balance and piecewise analytic algorithms are implemented. The integration step h of the equations of motion is chosen so that the particle crosses a difference-grid cell in several time steps

$$(1/2)\,\delta\bar{h} \leqslant |\,\mathbf{r}_{n+1} - \mathbf{r}_n\,| \leqslant 2\delta\bar{h},$$

where δ is a given parameter and $\bar{h}$ is the average value of the grid step.

When calculating beams in an external magnetic field, we arrive at a solution of the Cauchy problem for Poisson's equation. It appears to be important to use algorithms stable with respect to errors in the initial data. One such algorithm is based on Poisson's formula for a potential and the introduction of the operator that regularizes in the sense of A. N. Tikhonov [9.12]. The harmonic function at a point $(r,\ z)$ is defined approximately as

$$\varphi_\alpha(r,\ z) = \frac{R^2 - r^2 - z^2}{2\pi} \int_0^\pi \int_{-R}^{R} \frac{\tilde{\varphi}_\alpha(s)\,ds\,d\theta}{[R^2 + r^2 + z^2 - 2r\,(R^2 - s^2)^{1/2}\cos\theta - 2rs]^{3/2}}$$

$$= \int_{-R}^{R} K(r, z, s)\,\tilde{\varphi}_\alpha(s)\,ds, \qquad (9.60)$$

where R is the radius of a sphere enveloping the computational region and $\tilde{\varphi}_\alpha(s)$ is a "regularized" potential distribution on that sphere.

To find it we set up an integro-differential equation

$$\int_{-R}^{R} K(0, z, s)\,\tilde{\varphi}_\alpha(s)\,ds - \alpha[\tilde{\varphi}''_\alpha(s) - \tilde{\varphi}_\alpha(s)] = \varphi(0, z), \tag{9.61}$$

where α is the regularization parameter. In addition we impose the natural conditions $\tilde{\varphi}'_\alpha(-R) = \varphi'_\alpha(R) = 0$ on $\tilde{\varphi}_\alpha(s)$. If $\varphi(0, z)$ is given at M points of the axis, then we can obtain a system of linear algebraic equations in $\tilde{\varphi}_\alpha(s_i)$, $i = 1, 2, \ldots, M$ by using quadrature formulas and finite differences to approximate the left-hand side of (9.61). After solving the system (this can be done by Gauss elimination) it is then easy to calculate $\varphi_\alpha(r, z)$ by applying quadrature formulas to integral (9.60). The parameter α can be chosen such that the standard deviation of the calculated values $\varphi_\alpha(r, z_i)$ from the given ones $\varphi(0, z_i)$ be minimum.

Another algorithm is based on the use of a segment of the Scherzer series, which is an expansion of a harmonic function

$$\varphi(x, y) = \varphi(x, 0) - \frac{1}{2!}\frac{\partial^2\varphi(x, 0)}{\partial x^2}y^2 + \frac{1}{4!}\frac{\partial^4\varphi(x, 0)}{\partial x^4}y^4 - \ldots, \tag{9.62}$$

for the plane field and a similar relation

$$\varphi(r, z) = \sum_{\nu=0}^{\infty}[(-1)^\nu/(\nu!)^2](r/2)^{2\nu}\,\partial^{2\nu}\varphi(0, z)/\partial z^{2\nu} \tag{9.63}$$

for the axisymmetric case.

Expressions (9.62), (9.63) describe the solution of the Cauchy problem for the Laplace equation in a Cartesian and a cylindrical coordinate system respectively for the given values of φ and the zero normal derivative on the line of symmetry.

The approximation of the field on the axis (or some straight line parallel to the axis) is carried out using smoothing splines [9.13].

9.4. A COMPARATIVE MODULAR ANALYSIS OF METHODS FOR SOLVING PROBLEMS IN ELECTROPHYSICS

The solution of the boundary-value problems of mathematical physics and questions concerning the relevant algorithms and packages of application programs (PAP) may be presented in relation to several stages.

A. Preparation of initial data. This includes the description of the differential equation (or system of equations), its coefficients, the configuration of the computational region, the boundary conditions, and the initial data. Additionally variable parameters, their con-

straints and the objective function (the minimized functional) have to be specified for optimization problems. If the software provides for the use of different methods, then the necessary algorithms or their computational parameters (the type of grid, the number of nodes, the form of the grid equations, the method of solving them, etc.) must also be supplied. Finally, this stage may also include information about the content and form of the results or messages concerning the process of the computation to the user.

B. Discretization of the differential problem. At this stage the grid is constructed, the grid equations formed, the approximation of differential or integral equations and boundary conditions are specified, and the matrices and vectors of the resultant systems of algebraic equations are formed.

C. Solution of grid equations. In nonstationary problems the solution is reduced to a calculation of the grid functions for a sequence of time intervals, in stationary problems either iterative or direct methods of solving the algebraic equations, which are highly specific because of the band structure of the matrices that are formed, are used.

The current efficient approaches to the solution of grid problems are based on a combination of direct and iterative methods. Approaches like Schwartz's alternating method, where the computational region is partitioned into subregions while calculating in turn successive approximations to the solution in each of them are efficient. When uniform grids are used in simple subregions, fast Fourier transform or cyclic reduction methods are especially economical. Sequences of grids can also be useful. First an approximate solution is sought on an auxiliary sparse grid, then an interpolation is made into the nodes of the dense grid, and then the solution is refined on that grid. The volume of calculations can thus be substantially reduced and at the same time the accuracy increased by combining solutions on different grids (the Richardson extrapolation).

D. Processing and analysis of the results. The immediate result of the grid equations is a vast ocean of information. The user tends only to need a few values at separate points and on some lines, e.g. the values of extremes and the coordinates of the corresponding points, isolines, the vector fields of the gradients, and different functionals of the solution. It is extremely useful for the user can visualize the results and hence various graphical methods, e.g. graphs, tables, different number formats, headings and textual explanations. The use of a package is greatly enhanced if there is a flow of intermediate data such as messages concerning the progress of the computation, diagnostics of the errors made by the user and recommendations for their elimination, and estimates of computational error (if possible). All this requires special software, especially computer graphics, and the output peripherals must be such to ensure the efficiency and completeness of the analysis.

The basic modules are easy to discern from the above.

The first group should include the input modules. It is useful to have information about a problem in two forms. The first is designed to be easy for the user in describing the problem and the basic requirements are conciseness, visualization, and versatility. The way of describing the data may be called an input language.

The second representation of the initial data (we call it by convention an internal language) is module-oriented, and allows the modules implementing the algorithms to handle the data economically. The use of an internal language makes it possible to develop and modify the modules regardless of the external interface between the package and user. The module that translates the information from the input to the internal language is a translator whose complexity is determined by the level of the expressive means to be implemented by it.

The second group of modules constructs the grid. There is a vast literature on the topic and the algorithms capable of automatically constructing grid in two- and three-dimensional regions satisfying the basic necessary requirements. The output from these modules are the coordinates of grid nodes (appropriately ordered, which is also important), their positions relative to the boundary and their "neighbours" and other information necessary for third group of modules, which construct grid equations.

The modules that solve the grid equations carry a considerable computational load since they tend to recalculate the grid functions for a large number of iterations and/or time steps. These modules are most often simple in structure since the basic logical and geometrical operations are best completed in the preceding stages. The situation changes of course if large-scale arrays must be handled at each time or iteration step. Important decisions in such cases concern the sectionalization of the arrays and the organization of rational exchanges with the external memory.

In good packages there is a large group of service modules to process and output results, to control the user's work and the package itself, and to ensure the work of the calculator modules.

9.5. SOFTWARE SAPR REQUIREMENTS

We consider a technological array of the process of development, functioning and evolution of a software SAPR.

The starting point is the object to be designed. This may be an electrophysical device for which input data are given and the characteristics are specified. In the simplest case the problem is to define quantitatively the device's characteristics for given input conditions. For example, high voltage constructions mainly require the peak values of the electrostatic intensities for fixed potentials and

configurations of metallic electrodes with a known shape of insulation gap and known dielectric properties. But the main aim when constructing new electrophysical equipment is to find the best geometric and electrical parameters to ensure the necessary service properties. Often the geometry of the electrodes has to be found given some constraints on the overall dimensions and such that the intensity modulus does not exceed the critical value that may lead to a break-down. Thus we arrive at a synthesis or optimization problem at the basis of the design process.

The SAPR application software for electrophysical devices will contain some complex mathematical formulations and may number tens of thousands of statements, while the volume of data may run into hundreds of thousands of words and the calculations may have to be run several times. Thus there is a considerable expenditure of time and computer resources.

The development and operation of a SAPR software is a long process and particular attention must be paid to the methodology of constructing the programs, to their operation and supervision.

The industrial character of this sort of software must satisfy strict requirements which may be formulated as follows:

(a) completeness of the physical and mathematical models embraced by the structure of the developed programs;

(b) efficiently implemented algorithms, a measure of which being to a first approximation the amount of machine time required for a solution to a given accuracy. A better measure must take into account the cost of all the computer resources expended on a calculation and the effort that goes into the development of the programs;

(c) simple and easy operation: dynamic running of problems on the computer, visual and economical representation of source information and computational results, ease of access to the interface with the computer (dialogue or in batch operation, error and partial result messages), complete and rigorous documentation;

(d) reliable algorithms and programs: a high degree of debugging and safe operation, ensuring the required accuracy, the availability of various tests;

(e) extensible algorithm and program structures, the possibility of changing or extending the mathematical model without having to update the software substantially;

(f) software easily adaptable to changes in the configuration of the computer system.

These items are naturally part of the general topic of software quality and for an exhaustive treatment of these questions see [9.1].

An important point is the admissibility of different levels of using an application software package [9.14]-[9.17]. The most typical application should be to carry out designs using standard circuits.

On the other hand, the possibility of adding new components to enable numerical methods to be optimized or the class of physical and mathematical models to be extended is of considerable practical importance.

These features mean that the software must be a product which is divorced from its designer, that in the manuals must be good and, moreover, its structure will depend on the intended operation.

A SAPR that satisfies to some extent these requirements is in fact a package of programs, which is defined as a collection of programs compatible in terms of data structures and control, and united by the common functional purpose of solving a class of problems by a certain circle of users.

9.6. EXAMPLES OF PAP'S FOR THE AUTOMATIC DESIGN OF ELECTROPHYSICAL DEVICES

We now describe several application software packages implemented at the Laboratory for the Automatic Construction of Algorithms, the Computing Centre of the Siberian Division of the USSR Academy of Sciences within the framework of the software SAPR of electrophysics.

The ERA package is designed to calculate stationary high current relativistic beams in two-dimensional (plane or axisymmetric) fields.

The problem is to calculate the electrostatic field, the charged particle trajectories, and the charge and current density distribution. The problem is solved using the self-consistent formulation described by the equations in Sec. 9.1. The potential distribution is described by Poisson's equation in Cartesian or cylindrical coordinates. The computational region is bounded and may have practically any arbitrary boundary individual sections of which are represented by line segments or second-degree curves. On the various boundary segments boundary conditions of the first, second or third kind may be given. It is assumed that there are several subregions with different dielectric properties at whose interfaces the continuity condition of the potential and the jump condition of the normal derivative are established.

Several beams of particles with different charges and masses may be simulated in the computational region. The initial current densities may be either described by a given function of coordinates and velocities or given that the current is limited by the space charge. When particles move, the azimuthal component of an internal magnetic field and all the components of the external magnetic field given by values on the axis or plane of symmetry are taken into account. When the accelerating voltages are high relativistic effects are allowed for.

An important feature of the package is that it is oriented towards users who are engineers and hence there is a special method of describing complex boundary-value problems and the initial data for particles, the automation of the construction of the algorithms and the output of the computational results in alphanumeric and graphical form.

Poisson's equation is solved by the finite-difference methods described in Sec. 9.3 on rectangular grids with piecewise uniform mesh steps. The approximation is done on five-point or nine-point patterns. For near-boundary nodes approximations of the boundary conditions of all the above types, including those on curvilinear boundaries, are used.

A considerable proportion of the package is taken up by algorithms for automatically constructing the difference equations which include the time-consuming analysis of where a grid node is relative to the boundary, i.e. whether it is exterior or interior to the computational region, how far it is to the boundary along the coordinate lines, and what kind of boundary conditions there are on the sections adjoining the node. For near-boundary nodes, the necessary geometrical constructions (drawing normals to the boundary, calculating where they intersect the grid lines, interpolation, etc.) are carried out automatically allowing for boundary conditions with derivatives.

The way the coefficients of difference equations are represented and stored is essential for economic algorithms. In the ERA package, the computed coefficients and the right-hand sides are only stored for near-boundary nodes (6 numbers are stored for each node in the case of five-point patterns and 10 in the case of nine-point patterns). During the iterative solution of the difference equations, for regular nodes the coefficients are calculated from simple formulas via grid steps and for near-boundary nodes precomputed values are used. This allows the package to avoid a cumbersome geometric analysis at each iteration. Storing the difference coefficients for each node would reduce the time required for iteratively solving the equations but it would significantly increase the working storage required and thus would lead to severe constraints on the number of grid nodes.

The difference equations may be solved by any one of the algorithms in the package, which are pointwise overrelaxation (SOR), by overrelaxation with line sweeps (SLOR), or overrelaxation criss-cross sweeps (ADI). To simplify the logic, the iterations in every method are assumed to be implemented for every grid node, including the exterior ones. The type of each node is determined (the corresponding coding of the nodes is carried out during the geometrical analysis and the calculation of the coefficients). No calculation is done for the exterior nodes while successive approximations are calculated for the interior ones, so that given a homogeneous com-

putational scheme the boundary conditions for an arbitrary form of boundary, including a multiply connected one, are allowed for owing to the difference coefficients for the near-boundary nodes.

The charged-particle trajectories are calculated by integrating the equations of motion in momentum form.

The current and space charge densities are calculated as described in Sec. 9.3, and calculating "exterior" space charge iterations. The external magnetic field is determined using spline approximations from the values H_z on the symmetry line.

Self-consistent problems may be solved on a sequence of nested grids, in which case first a rough approximation is found on an auxiliary sparse grid (with four fewer nodes than those on the original grid) and then the results are refined on the principal grid.

The most complex source information concerns the description of the boundary-value problem. To specify it the input language PG2M [9.16] is used. This is capable of representing an ungeometrical boundary and boundary conditions in a visual and economical form. The translator from that language in the ERA package translates character texts into the number arrays subsequently processed by the computing modules.

The second group of input data relates to the "trajectory" part of the package. To calculate the motion of a beam of charged particles it is necessary to specify such parameters as the geometry of the surface where the beam enters the computational region (the initial front of the beam), the energy and angular responses of the particles, the modes of operation of the emitter, the components of the external magnetic field, the information on the output of trajectories, and some parameters used in numerical algorithms. To describe the information ERA uses a modular analysis to break the input data into information modules, each having parameters of the same semantic type. To process a module a FORTRAN subroutine is developed. Essentially the subroutine takes in the information in the module via its parameters, processes it, and transfers it into a standard file which consists of a number sequence "intelligible" to the computing modules. The standard file is set to default values corresponding to the most frequently used parameter values.

The PAP is a collection of data consistent FORTRAN, ALGOL or assembler subroutines (modules) executed under the control of the package's master routine and the DUBNA monitoring system on a BESM-6 computer. The modules of the package are:

PG2M, the translator from the input language for the boundary problem,

GEOM, the subroutines to process and calculate the geometric information, grid nodes, and difference equation coefficients,

ITER, the module to solve the difference equations iteratively,

BEAM, the module for computing trajectories, and current, space-charge and magnetic field densities,

SERV, the service module for outputting the computational results in alphanumeric or graphical form.

When the most laborious self-consistent problems are solved the main computational load is carried by the ITER and BEAM modules. In this case the control module calls them in turn and iterates "exterior" space-charge iterations with relaxation.

For a more detailed description of the ERA software see [9.5].

The ESTAMP PAP software package is designed to calculate two-dimensional magnetostatic fields (plane or axially symmetric) allowing for the nonlinear properties of isotropic ferromagnetic materials.

The problem reduces to the solution of the boundary-value problem in a bounded region for a vector potential component (in the axisymmetric case this is the azimuthal component). The boundary of the region may be multiply connected and consist of straight lines or second degree curves. Arbitrary boundary conditions of the first, second or third kind are assumed on the individual sections of the boundary. It is assumed that the region has an arbitrary number of coils with given current densities and arbitrary number of magnetocircuits with different configurations, the latter consisting of several types of ferromagnetic materials whose magnetic properties are given by the positions of their B/H characteristics.

These boundary-value problems are solved by the finite-element method. The computational region is triangulated and the original problem is approximated in the space of piecewise linear functions, which leads to a system of nonlinear seven-point grid equations for the values of vector potential components.

Nonlinear equations are solved by the SOR-Newton iteration method. To enhance the efficiency a precomputation on a sparse grid with twice the mesh spacings is used, after which the results are interpolated into the nodes of the principal grid and the resulting approximation is reiterated. Allowing for the nonlinear properties of ferromagnetics the B/H characteristics are approximated from given points by third-degree splines. To obtain a good initial estimate of a nonlinear problem first the linear equations corresponding to the dielectric permittivity of the ferromagnetic (it is chosen from a priori physical considerations) is solved.

To calculate the fields of magnetic lenses ESTAMP also implements a version of the subregion iteration. First the field in the nonferromagnetic medium is calculated, with the Neumann homogeneous condition imposed on the boundary with the "iron". The grid equations are then solved in the "iron" with the Dirichlet boundary conditions obtained from the previous solution. We thus obtain a first approximation, and then the iterations are continued in a similar way alter-

nating the type of the boundary conditions used. When the field in an air + current coil subregion is calculated a Neumann condition with the value of the normal derivative determined from the previous approximation in the "iron" is imposed each time on the boundary of iterated subregions. For some types of magnetic lenses (when the values of ν are much less than unity) even the first iteration yields a reasonably good approximation.

In order to calculate the magnetic field of electron-optical systems, as a rule, only the intensity component on the line of symmetry need be precisely computed. The solution is then refined by focussing. A rectangular subregion (a "window") that includes the line segment of interest is chosen and the boundary conditions obtained on its boundary from the solution of the basic grid equations are taken. Then a dense uniform rectangular grid on which a refined solution can be calculated quickly is constructed in the "window".

The source information of the ESTAMP PAP includes a description of the boundary-value problem, the specification of the ways of triangulating the region, data on the B/H characteristic of the ferromagnetics (if they are not available in the package's data base) and directions for the mode of calculation and the output of the required results.

The basic data are given in PG2M.

In addition the input representation specifies the values of relaxation parameters, the required accuracy and control characters for the mode of calculation (if necessary, partial quantities are recorded on disc, or the results output on the graph plotter).

The ESTAMP is a collection of FORTRAN data consistent subroutines executed under the control of a single master routine and the DUBNA monitoring system on a BESM-6 computer. The exception is the translator from the input language which is implemented in extended ALGOL. The modules operate thus:

The data is input by PG2M, translated into the internal representation, and diagnostics of the user's errors are issued.

The grid is constructed and its quality controlled with messages issued if there are irregularities, angles are too small, etc.

The coefficients of the spline approximation to the B/H characteristic are calculated.

The grid equations are iteratively solved for vector potential components (this stage of calculation is the most laborious and interesting in terms of the inherent logic). The calculations may be carried out either on a sequence of grids or using subregion iterations or with different values of relaxation parameters or by a combination of any of these ways. The solution is refined (if necessary) on the line of symmetry by focussing.

The results are processed and the necessary physical characteristics output in an alphanumeric or graphic form.

The EFIR PAP was designed to calculating two-dimensional (plane or axially symmetric) electrostatic fields using integral equations for the potential of a simple layer (see [9.2]). In addition, this package makes it possible to calculate disturbances in the fields of axisymmetric electron-optical systems (EOS) and to optimize them by minimizing a given functional of the solution while varying the potentials or geometric parameters of electrodes for specified linear or nonlinear constraints.

The Laplace equation is considered in either Cartesian or cylindrical coordinates. The solution must be formed in a bounded or unbounded region with a boundary (possibly multiply connected) made up of line segments or portions of second-degree curves. It is assumed that the region may have "internal" boundaries on which the continuity condition of the solution and that of the jump of its normal derivative are given. Boundary conditions of the first, second or third kind may be given on different sections of the boundary.

The problem of optimizing electrostatic fields is considered in a formulation similar to the one considered in Sec. 9.3.

The EFIR package allows the following types of geometric variations: a shift of a section of the boundary along a given vector, a rotation of a section of the boundary about a fixed point, a longitudinal extension (compression) of a line segment with a given fixed point, a change in the radius of a circle segment with one fixed point on the boundary. In addition the related variations are considered, as when a rotation of one segment results in a shift or compression of the others, etc.

We first discuss briefly the simplest problem, that of calculating a solution of the Laplace equation for given boundary conditions.

The desired harmonic function is represented as the potential of a simple layer. For the surface density a system of integral equations is formed from the boundary conditions. If there are singularities on the boundary the asymptotic behaviour of the density is taken into account.

The system of algebraic equations is solved by orthogonalizing it with the stability of calculations automatically controlled and an a posteriori estimation of the error of the result.

After the surface density has been estimated the desired potential is determined from the approximation of the original integral representation using the B-splines and quadrature formulas.

The derivatives of the potential are calculated from the expressions obtained by differentiating the integral representation of the potential.

The calculation of nonaxisymmetric disturbances in the electron-optical systems is calculated using integral equations for individual harmonics.

The source information about the boundary-value problem is spe-

cified in the EFIR package in the IC input language (input complex) which is a development of the languages used in the ERA and ESTAMP.

The EFIR package was written entirely (including the translator from the input language) in FORTRAN and implemented on both a BESM-6 and an ES computer.

The structure of the modules divide functionally thus: processing the source character information and its initial analysis with error diagnostics, controlling the computation, forming the systems of linear algebraic equations (a laborious and logically difficult task), solving the system, solving the problem of minimization with constraints, and processing and outputting results.

The control module uses an analysis of the input data to determine the form of the assignment, that is whether only the boundary-value problem is to be solved or to calculate nonaxisymmetric disturbances or to solve an optimization problem. It also carries out the necessary exchanges with the external memory if communications are necessary between different computational sessions.

The matrices of the algebraic systems are formed in a uniform manner for different kernels and right-hand sides of the integral equations corresponding to the different types of problems.

The output module ensures the formation of visual alphanumeric lists, tables, and sketches.

The EDIP PAP is for calculating the fundamental or first few natural frequencies and the corresponding harmonics of electrodynamic systems and to determine their basic physical characteristics. The EDIP package also allows two-dimensional (plane and axially symmetric) fields described by boundary-value problems to be calculated.

Two-dimensional free-oscillation and wave problems which reduce to the eigenvalue problem for the Laplace equation in a bounded region Ω with a piecewise smooth boundary Γ are considered. In a physical formulation, with $\alpha = 0$ (in Cartesian coordinates) the computational region Ω is a cross-section of a regular waveguide, and with $\alpha = 1$ it is part of the meridional plane of an axisymmetric resonator. The Dirichlet or Neumann boundary conditions for an oscillation mode are imposed under the assumption of ideal wall conduction.

From the solution of the original problem various physical characteristics of the electrodynamic systems are determined, viz. wavelengths, the coefficients of amplification of fields, characteristic impedances, etc.

To solve the eigenvalue problem or to solve the boundary-value problem of Poisson's equation the finite-difference methods described in Sec. 9.3 are used. The original differential equation is approximated by a system of five-point difference equations, whose coefficients in the neighbourhood of singular points are calculated automatically

allowing for the asymptotic behaviour of the solution. In the calculations a sequence of nested densed rectangular piecewise constant grids is used.

To refine the difference solutions Richardson's extrapolation is used in which a combination of two solutions on grids makes it possible to obtain an error up to $O(h^4)$. Naturally, this increases the economy of the calculations since a close solution from the grid Ω_h is taken as initial estimate when solving a problem on a grid $\Omega_{h/2}$.

The algebraic eigenvalue problem for a system of difference equations is solved by various iterative algorithms. If only a minimum eigenvalue and the corresponding eigenvector are to be calculated, then the Rhue algorithm, which is based on successive overrelaxation with an automatic choice of a "nearly optimal" relaxation parameter, turns out to be the most reliable and economical method.

To calculate several fundamental natural frequencies and their harmonics, the subspace iteration method is efficient.

Using the resultant harmonics obtained the components of the electric and magnetic fields can be calculated by approximating the above formulas by finite differences with a second-degree error.

A feature of the EDIP package is the availability of two independent ways of specifying input data, each constructed on different principles. The first approach is to describe the information with the aid of a character text in a natural form in a way similar to that used in the ERA, ESTAMP, and EFIR packages.

The other approach is to create a set of auxiliary procedures for each package, each virtually a descriptor of a boundary section with an index and type of boundary conditions assigned to it.

The user can instruct the package to output to the printer or graph plotter tables, graphs, or sketches with the required textual comments. These may be the values of the harmonics or those of the intensity components at the grid nodes, the frequencies, wavelengths, characteristic impedances of the electrodynamic system, etc. If there are user errors in the input data diagnostic messages are issued.

The algorithms for designing electrodynamic systems have been implemented as subroutines in FORTRAN, ALGOL-DDR, and MADLEN autocode for the DUBNA monitoring system for a BESM-6.

The EDIP modules are functionally those for processing the input data (in particular, translating the character information into the internal language), automating the construction of the difference equations and forming the system's matrix, solving the algebraic equations or the eigenvalue problem, and servicing auxiliary operations. The computation is organized by the executive routine which calls the modules turn and interfaces them, accesses external memory, and outputs results as described by the input data.

References

9.1. V. P. Ilyin, *Numerical Methods of Solving Problems of Electrooptics* (Novosibirsk: Nauka, 1974) (in Russian).

9.2. V. P. Ilyin and V. A. Kateshov, *Avtometriya* (Novosibirsk), No. 4: 67-74 (1982) (in Russian).

9.3. A. L. Urvantsev, in: *Numerical Methods of Solving Problems of Electron Optics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 77-83 (in Russian).

9.4. V. P. Ilyin and A. V. Gavrilin, *Avtometriya* (Novosibirsk), No. 4: 74-79 (1982) (in Russian).

9.5. N. I. Gorbenko *et al.*, in: *Numerical Methods of Solving Problems of Electron Optics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 34-60 (in Russian).

9.6. B. I. Golubtsov and V. P. Ilyin, *Sibirsk. Matemat. Zhurnal* (Novosibirsk) XIX, No. 3: 566-570 (1978) (in Russian).

9.7. G. I. Marchuk, *Methods of Computational Mathematics*, 2nd ed., rev. and enl. (Moscow: Nauka, 1980) (in Russian).

9.8. S. A. Sander, in: *Sparse Matrix Computations* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1981): 123-133 (in Russian).

9.9. S. A. Sander, *Modification of the Schwartz Algorithm for Solving Grid Boundary-Value Problems in Regions Made Up of Rectangles and Rectangular Parallelepipeds*, Preprint No. 83 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1981) (in Russian).

9.10. A. V. Gavrilin and N. I. Gorbenko, in: *Numerical Methods of Solving Problems of Electron Optics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 99-107 (in Russian).

9.11. A. V. Gavrilin and V. P. Ilyin, *On an Iterative Method for Solving the Partial Eigenvalue Problem*, Preprint No. 85 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1981) (in Russian).

9.12. G. S. Popova and M. V. Urev, in: *Numerical Methods of Solving Problems of Electron Optics* (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979): 89-98 (in Russian).

9.13. M. V. Urev, *Radiotekhnika i Elektronika* 28, No. 4: 772-779 (1983) (in Russian).

9.14. S. P. Gololobova, V. P. Ilyin, and G. A. Plotnikova, *The MOPR System of Modular Programming, A General Description*, Preprint No. 261 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1980) (in Russian).

9.15. A. P. Ershov and V. P. Ilyin, in: *Application Software Packages, Algorithms and Algorithmic Languages* (Moscow: Nauka, 1982) (in Russian).

9.16. V. P. Ilyin, *Avtometriya* (Novosibirsk), No. 3: 77-80 (1980) (in Russian).

9.17. V. P. Ilyin, in: *Development of Application Software Packages* (Novosibirsk: Nauka, 1982): 113-129 (in Russian).

9.18. V. P. Ilyin and V. A. Kateshov, *Automation of the Solution of Two-Dimensional Boundary-Value Problems*, Preprint No. 173 (Novosibirsk: Vychisl. Tsentr SO AN SSSR, 1979) (in Russian).

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

Printed in the Union of So

ANOTHER TITLE IN THE SERIES

MACROSCOPIC THEORIES OF MATTER AND FIELDS:

A THERMODYNAMIC APPROACH

Edited by L.I. Sedov, Mem. USSR Acad. Sc.

This is a collection of articles by Soviet scientists on current issues of building macroscopic models of matter and fields. Based on thermodynamics concepts the papers develop general variational techniques of modeling material continuous media and fields allowing for their interactions in reversible and irreversible processes.

The book is intended for researchers, engineers, graduate and postgraduate students interested in the mechanics of continuous media.